arXiv:2109.09203

Topology, Convergence, and Reconstruction of Predictive States

Samuel P. Loomis* and James P. Crutchfield†
Complexity Sciences Center and Department of Physics and Astronomy, University of California at Davis, One Shields Avenue, Davis, CA 95616

(Dated: July 4, 2022)

Predictive equivalence in discrete stochastic processes has been applied with great success to identify randomness and structure in statistical physics and chaotic dynamical systems and to inferring hidden Markov models. We examine the conditions under which predictive states can be reliably reconstructed from time-series data, showing that convergence of predictive states can be achieved from empirical samples in the weak topology of measures. Moreover, predictive states may be represented in Hilbert spaces that replicate the weak topology. We mathematically explain how these representations are particularly beneficial when reconstructing high-memory processes and connect them to reproducing kernel Hilbert spaces.

Keywords: stochastic process, symbolic dynamics, dynamical systems, measure theory, weak topology

I. INTRODUCTION

With an accurate model in hand, an observer can leverage their knowledge of a system's history to predict its future behavior. For stochastic processes—distributions over time-series data—the task of predicting future behavior from past observations and the associated resource constraints this task imposes on an observer have been studied under the physics of *computational mechanics* [1]. This subfield of statistical mechanics focuses on the intrinsic information-processing embedded in natural systems.

Its chief insight is the concept of the predictive (or causal) state. A process' predictive states play a dual role. On the one hand, to accurately predict a process' future behavior they are the key objects that an observer must be capable of reproducing in their model. On the other hand, the predictive states and their dynamics are central to understanding the intrinsic, model-independent properties of the process itself [1].

The concept of predictive states has found use in numerous settings, such as classical and quantum thermodynamics [2–4], quantum information and computing [5–8], condensed matter [9–11], dynamical systems [12], cellular automata [13], and model inference [14–20]. Additionally, in the setting of processes generated by finite-state, discrete-output hidden Markov models (HMMs) and generalized hidden Markov models (GHMMs), a deep mathematical theory of predictive states is now available [1, 21–25].

Despite their broad utility, a mathematically rigorous definition of predictive states is needed that is applicable and useful for even more general stochastic processes. Here, we have in mind large-memory processes whose long-time correlations cannot be finitely represented by HMMs or GHMMs and processes whose outputs may span a continuous domain in time and space.

The following takes the next major step towards a rigorous and mathematically general definition of predictive states, developing an approach which uses the tools of functional analysis to better understand the topology and structure of predictive states, and which applies to all processes whose observations are temporally discrete but may otherwise be either discretely- or continuously-valued.

It has been previously noted [21] that predictive states are always well-defined and can be convergently approximated from empirical observations for any discretely-valued stationary and ergodic process. We extend and deepen this result, relating this convergence to the topology of sequences and applying it to continuously-valued stochastic processes. Next, we expand on recent work on Hilbert space embeddings of predictive states [20, 26–28], and demonstrate that such embeddings always exist and discussing their implications for predictive-state geometry and topology. Last, we explore the implications of our results for empirically reconstructing predictive states via reproducing kernel Hilbert spaces, particularly through the addition of new terms in the asymptotic convergence bounds.

II. ASSUMPTIONS AND PRELIMINARIES

A. Stochastic processes

We begin by laying out a series of definitions and identifying the assumptions made. We draw from the combined

* sloomis@ucdavis.edu
† chaos@ucdavis.edu

literature of measures, stochastic processes, and symbolic dynamics [29, 30].

A *stochastic process* is typically defined as a function-valued random variable $X : \Omega \to \mathcal{X}^{\mathcal{T}}$, where (Ω, Σ, μ) is a measure space, $\mathcal{T}$ is a set of temporal indices (perhaps the real line, perhaps a discrete set), and $\mathcal{X}$ is a set of possible observations (also potentially real or discrete in nature). We take the sample space Ω to be the set $\mathcal{X}^{\mathcal{T}}$ and X to be the identity. In this way, a stochastic process is identified solely with the measure μ over $\Omega = \mathcal{X}^{\mathcal{T}}$.

When $\mathcal{T}$ is $\mathbb{Z}$, we say the process is *discrete-time*; when it is $\mathbb{R}$ we say *continuous-time*. Unless specified otherwise we assume discrete-time, later treating continuous-time as an extension of the discrete case. In discrete time, it is convenient to write $X(t)$ as an indexed sequence (x_t), where each x_t is an element of $\mathcal{X}$. When $\mathcal{X}$ is a discrete finite set, we say that the process is *discrete-observation*; by *continuous-observation* we typically mean the case where $\mathcal{X}$ is an interval in $\mathbb{R}$ or a Cartesian product of intervals in $\mathbb{R}^d$. These are the only cases we consider rigorously. That said, we believe they are sufficient for many practical purposes or, at least, not too cumbersome to extend if necessary; in Appendix I we discuss possible extensions to noncompact $\mathcal{X}$ and continuous time.

The temporal *shift operator* $\tau : \mathcal{X}^{\mathcal{T}} \to \mathcal{X}^{\mathcal{T}}$ simply translates $t \mapsto t+1$: $(\tau X)(t) = X(t+1)$. It also acts on measures of $\mathcal{X}^{\mathcal{T}}$: $(\tau\mu)(A) = \mu(\tau^{-1}A)$. A stochastic process paired with the shift operator—$(\mathcal{X}^{\mathcal{T}}, \Sigma, \mu, \tau)$—becomes a dynamical system and is *stationary* if $\tau\mu = \mu$. It is further considered *ergodic* if, for all shift-invariant sets $\mathcal{I} \subseteq \mathcal{X}^{\mathcal{T}}$, either $\mu(\mathcal{I}) = 1$ or $\mu(\mathcal{I}) = 0$. Here, we assume all processes are both stationary and ergodic.

A consequence of ergodicity, and an equivalent restatement of the property, is given by the following convergent limit for every measurable function $f : \mathcal{X}^{\mathbb{Z}} \to \mathbb{R}$ and μ-almost every $X \in \mathcal{X}^{\mathbb{Z}}$:

$$\lim_{N\to\infty} \frac{1}{N} \sum_{t=0}^{N-1} f(\tau^t X) = \int f(X) d\mu(X)$$

That is, temporal Monte Carlo averages will converge to averages over the stationary measure μ. The convergence rate, and its dependence on X, will vary from process to process and has no standard form since we are making no additional assumptions beyond ergodicity.

Examples of stationary and ergodic processes include Markov chains and hidden Markov chains; however, the class is much broader than either of these, including also renewal and hidden semi-Markov processes and processes generated by probabilistic stack automata.

If $\mathcal{X}$ is discrete, then the measurable sets of $\mathcal{X}^{\mathbb{Z}}$ are generated by the *cylinder sets*:

$$U_{t,w} := \{ X : x_{t+1} \dots x_{t+\ell} = w \} \ ,$$

where $w \in \mathcal{X}^\ell$ is a *word* of length ℓ. For a stationary process, the *word probabilities*:

$$\Pr_\mu (x_1 \dots x_\ell) := \mu (U_{0,x_1 \dots x_\ell})$$

are sufficient to uniquely define the measure μ.

In the continuous-observation case, the issue is more subtle. A cylinder set instead takes the form:

$$U_{t,I_1 \dots I_\ell} := \{ X : x_{t+1} \in I_1, \dots, x_{t+\ell} \in I_\ell \} \ ,$$

where each I_t is an interval in $\mathcal{X}$ (or product of intervals, if $\mathcal{X} \subseteq \mathbb{R}^d$). This does not lend itself well to expressing simple word probabilities; however, we can write the cylinder probabilities:

$$\Pr_\mu (I_1 \dots I_\ell) := \mu (U_{0,I_1 \dots I_\ell})$$

Additionally we can define the *word measures* $\mu|_\ell$ by restricting μ to the set $\mathcal{X}^\ell$ describing the first ℓ values (so that $\mu|_\ell(I_1 \times \cdots \times I_\ell) = \Pr_\mu (I_1 \dots I_\ell)$). By choosing intervals of rational dimension centered on rational-valued points, we can fully characterize any measure using only a countable number of probabilities $\Pr_\mu (I_1 \dots I_\ell)$—a consequence of Carathéodory extension.

B. Topology, continuity, and convergence on $\mathcal{X}^{\mathbb{N}}$

A central feature of our result on predictive states is that they converge in distribution as more information from the past is provided. Convergence in distribution is defined in terms of continuous functions. Namely, a sequence of measures μ_k over $\mathcal{X}^{\mathbb{N}}$ converges to a measure μ in distribution if, for every continuous function $f : \mathcal{X}^{\mathbb{N}} \to \mathbb{R}$,

$$\lim_{k\to\infty} \int_{\mathcal{X}^{\mathbb{N}}} f(\overrightarrow{x}) d\mu_k(\overrightarrow{x}) = \int_{\mathcal{X}^{\mathbb{N}}} f(\overrightarrow{x}) d\mu(\overrightarrow{x})$$

To relate this definition of convergence to our own intuitions of stochastic processes, we must have a better understanding of continuity in sequence-space.

The definition of continuity depends on the product topology, whose neighborhoods are cylinder sets. For the discrete case, a simple rendering of the definition of continuity is this: a function f is continuous if, for every $\overrightarrow{x} \in \mathcal{X}$ and some small number $\epsilon > 0$, there is a sufficiently large time t such that $|f(\overrightarrow{y}) - f(\overrightarrow{x})| < \epsilon$ when-

ever $y_1 \ldots y_t = x_1 \ldots x_t$. In other words, if two sequences match sufficiently far into the future, then their function values will be arbitrarily close.

Another feature of continuity on $\mathcal{X}^{\mathbb{N}}$ comes to us by virtue of the compactness of the space. Since $\mathcal{X}$ is always assumed to be compact (by virtue of being either a finite set or a bounded region of $\mathbb{R}^d$), $\mathcal{X}^{\mathbb{N}}$ is compact as well, by the Tychonoff theorem. Then the Heine-Cantor theorem asserts that any continuous function on $\mathcal{X}^{\mathbb{N}}$ is *uniformly* continuous. This means that we can in fact strengthen our definition of continuity: for any small $\epsilon > 0$, there is a single time $t > 0$ after which it is guaranteed that $|f(\overrightarrow{y}) - f(\overrightarrow{x})| < \epsilon$ for any two $\overrightarrow{x}$ and $\overrightarrow{y}$ that match on the first t symbols: $y_1 \ldots y_t = x_1 \ldots x_t$. In other words, convergence occurs at (at most) a uniform rate at every point; there are no "straggler points" that take an arbitrarily long time to converge compared to other points.

For the following statement, we call a measure μ *full* if it assigns positive measure to every cylinder set.

Proposition 1 (Continuity via word averages)**.** *A function $f : \mathcal{X}^{\mathbb{N}} \to \mathbb{R}$ is continuous if and only if the functions:*

$$f_{\mu,\ell}(x_1 \ldots x_\ell) = \frac{\int_{U_{0,x_1 \ldots x_\ell}} f(\overrightarrow{x}) d\mu(\overrightarrow{x})}{\mu(U_{0,x_1 \ldots x_\ell})}$$

are continuous on $\mathcal{X}^\ell$ and converge to $f(\overrightarrow{x})$ uniformly over $\overrightarrow{x}$ and for every full measure μ, as $\ell \to \infty$.

Proof. *Suppose f is continuous; then it is uniformly continuous, and so for every $\epsilon > 0$ there is a t so that, for every $\overrightarrow{x}$ and μ, $|f_{\mu,\ell}(x_1 \ldots x_\ell) - f(\overrightarrow{x})| < \epsilon$. This follows since f will be close to $f(\overrightarrow{x})$ on the cylinder set being averaged over and so the average will be close. Further, continuity of $f_{\mu,\ell}$ will be inherited from the continuity of f. Thus, the forward implication is true.*

For the converse, consider the fact that $f_{\mu,\ell}$ can be extended to a function on $\mathcal{X}^{\mathbb{N}}$ as $f_{\mu,\ell}(\overrightarrow{x}) = f_{\mu,\ell}(x_1 \ldots x_\ell)$. Each of these extensions is necessarily continuous on $\mathcal{X}^{\mathbb{N}}$. Since they converge uniformly to f, f must be continuous by virtue of the Uniform Limit Theorem. (The latter states that a uniform convergence of continuous functions results in a continuous function.)

Let us keep in mind that if $\mathcal{X}$ is discrete, then the requirement that $f_{\mu,\ell}$ is continuous is trivial.

We can now demonstrate that convergence-in-distribution is equivalent to convergence over word distributions. We state the full result and then explain the implications afterward.

Proposition 2 (Equivalence of convergence-over-words and convergence-in-distribution)**.** *Let μ_k be a sequence of measures on $\mathcal{X}^{\mathbb{N}}$ and let μ be a measure over the same. Then $\mu_k \to \mu$ in distribution if and only if $\mu_k|_\ell \to \mu|_\ell$ in distribution for every ℓ, where $\nu|_\ell$ is the projection of the measure ν to $\mathcal{X}^\ell$.*

Proof. *By virtue of Prop. 1, on the one hand, for any continuous function f for all measures ν:*

$$\lim_{\ell \to \infty} \int f_{\nu,\ell}(x_{1:\ell}) d\left(\nu|_\ell\right)(x_{1:\ell}) = \int f(\overrightarrow{x}) d\nu(\overrightarrow{x}) \ .$$

If we replace ν with μ_k and μ, respectively, then convergence in distribution has the form:

$$\lim_{k \to \infty} \lim_{\ell \to \infty} \int f_{\mu_k,\ell}(x_{1:\ell}) d\left(\mu_k|_\ell\right)(x_{1:\ell}) = \lim_{k \to \infty} \int f(\overrightarrow{x}) d\mu_k(\overrightarrow{x}) = \int f(\overrightarrow{x}) d\mu(\overrightarrow{x})$$

for all continuous f.

On the other hand, convergence of $\mu_k|_\ell \to \mu|_\ell$ for all ℓ takes the form of the requirement:

$$\int f(\overrightarrow{x}) d\mu(\overrightarrow{x}) = \lim_{\ell \to \infty} \int f_{\mu,\ell}(x_{1:\ell}) d\left(\mu|_\ell\right)(x_{1:\ell}) = \lim_{\ell \to \infty} \lim_{k \to \infty} \int f_{\mu_k,\ell}(x_{1:\ell}) d\left(\mu_k|_\ell\right)(x_{1:\ell})$$

for all continuous f.

Equivalence of these two convergences then boils down to the interchange of limits $\lim_{\ell \to \infty} \lim_{k \to \infty} \leftrightarrow \lim_{k \to \infty} \lim_{\ell \to \infty}$. By the Moore-Osgoode theorem, this interchange is in fact valid whenever the limit:

$$\lim_{\ell \to \infty} \int f_{\mu_k,\ell}(x_{1:\ell}) d\left(\mu_k|_\ell\right)(x_{1:\ell}) = \int f(\overrightarrow{x}) d\mu_k(\overrightarrow{x})$$

is uniformly convergent over all k. This is guaranteed by Prop. 1, though, and so the two forms of convergence are equivalent.

This proposition guarantees that to demonstrate convergence in distribution, it is sufficient that the measures converge on their marginalizations to finite words. For discrete $\mathcal{X}$, this means that:

$$\lim_{k \to \infty} \Pr_{\mu_k}(w) = \Pr_{\mu}(w) \ ,$$

for all w, is equivalent to convergence in distribution. This is extremely convenient, as word probabilities are perhaps the most intuitive way to interact with the measure.

For the case of $\mathcal{X} \subset \mathbb{R}$, the situation is more subtle. The Portmanteau theorem [29] states that convergence in distribution is equivalent to a very weak *bounded* con-

vergence over open sets. In our case, this means that:

$$\liminf_{k\to\infty} \Pr_{\mu_k}(I_1 \ldots I_\ell) \geq \Pr_\mu(I_1 \ldots I_\ell) \ ,$$

for all open neighborhoods $I_1 \times \cdots \times I_\ell$ of any length, is equivalent to convergence in distribution. In fact, what we prove for predictive states is a somewhat stronger form of convergence than this. The latter maintains the equality at each ℓ. This is still not nearly as strong, though, as other forms of measure convergence and, in most practical cases, it is equivalent to convergence in distribution.

III. PREDICTIVE STATES

Each element $X \in \mathcal{X}^{\mathbb{Z}}$ can be decomposed from a bidirectional infinite sequence to a pair of unidirectional infinite sequences in $\mathcal{X}^{\mathbb{N}} \times \mathcal{X}^{\mathbb{N}}$, by the transformation $\ldots x_{-1}x_0x_1\cdots \mapsto (x_0x_{-1}\ldots, x_1x_2\ldots)$. The first sequence in this pair we call the *past* $\overleftarrow{X}$ and the second we call the *future* $\overrightarrow{X}$. From this perspective, a stochastic process is a bipartite measure over pasts and futures. The intuitive definition of a *predictive state* is as a measure over future sequences that arises from conditioning on past sequences. Heuristically, $\Pr_\mu\left(\overrightarrow{X} \,\middle|\, \overleftarrow{X} = x_0x_{-1}\ldots \right)$ represents the "predictive state" associated with past $x_0x_{-1}\ldots$.

Conditioning of measures is nuanced, especially when the involved sample spaces are uncountably infinite [31]. Of the many perspectives that define a conditional measure, the most practical and intuitive is that a conditional measure is a ratio of likelihoods—and, in the continuous case, a limit of such ratios. However, determining the manner in which this limit must be taken is rarely trivial.

The following considers first the case of discrete observations, where the matter is relatively straightforward. Then we examine the case of continuous observations, reviewing the previous literature on the nuances of this domain and extending its results for our present purposes. As we will see, in either case, the intuition of predictive states can be born out in a rigorous and elegant manner for any stochastic process satisfying the assumptions heretofore mentioned.

A. Discrete observations

Let $\overleftarrow{\mu}$ denote the projection of μ to pasts. (We define this in further detail below.) We begin with the following result, first noted by Upper [21]:

Theorem 1. *For all measures μ on $\mathcal{X}^{\mathbb{Z}}$, all $\ell \in \mathbb{N}$, all $w = x_1 \ldots x_\ell \in \mathcal{X}^\ell$, and $\overleftarrow{\mu}$-almost all pasts $\overleftarrow{X}$, where $\mathcal{X}$ is a finite set, the limit:*

$$\Pr_\mu\left(w \,\middle|\, \overleftarrow{X} \right) := \lim_{k\to\infty} \frac{\Pr_\mu(x_{-k}\ldots x_0x_1\ldots x_\ell)}{\Pr_\mu(x_{-k}\ldots x_0)} \quad (1)$$

is convergent.

In the discrete case, there are only a countable number of words w. Thus, if for each w the set of converging $\overleftarrow{X}$'s is measure-one, then the intersection of these sets is also measure one. That is, Eq. (1) converges for all w for μ-almost all $\overleftarrow{X}$. The convergent word probabilities define a measure $\epsilon[\overleftarrow{X}] \in \mathbb{M}(\mathcal{X}^{\mathbb{N}})$ over future sequences, uniquely determined by the requirement $\epsilon[\overleftarrow{X}](U_{0,w}) = \Pr_\mu\left(w \,\middle|\, \overleftarrow{X} \right)$. Combining this observation with Prop. 2 leads to the corollary:

Corollary 1. *For all measures μ on $\mathcal{X}^{\mathbb{Z}}$ and μ-almost every past $\overleftarrow{X} \in \mathcal{X}^{\mathbb{N}}$, the measures $\eta_\ell[\overleftarrow{X}]$ defined by:*

$$\eta_\ell[\overleftarrow{X}](U_{0,w}) = \Pr_\mu(w \mid x_{-\ell+1}\ldots x_0) \quad (2)$$

converge to $\epsilon[\overleftarrow{X}]$ in distribution:

$$\eta_\ell[\overleftarrow{X}] \to \epsilon[\overleftarrow{X}] \ , \quad (3)$$

as $\ell \to \infty$.

This $\epsilon[\overleftarrow{X}]$ is the *predictive state* of $\overleftarrow{X}$ and the function $\epsilon : \mathcal{X}^{\mathbb{N}} \to \mathbb{M}(\mathcal{X}^{\mathbb{N}})$, the *prediction mapping*. Corollary 1 is an important extension of Theorem 1, as it provides the topological context for understanding the convergence of predictive states. And, this is useful when we examine the relation to reproducing kernel Hilbert spaces.

We present an alternative proof to that used by Upper, though. This sets the stage for our proof in the continuous-observation case. Our strategy consists in redefining the problem.

The Eq. (1) limit can be recast as a *likelihood ratio*. The convergence of likelihood ratios is itself closely related to the theory of Radon-Nikodym derivatives between measures. Specifically, the Radon-Nikodym derivative can be computed as a convergence of likelihood ratios if (i) that convergence is taken over a particular class of neighborhoods, called a *differentiation basis*, and (ii) that basis has a property called the *Vitali property*. We define these concepts below and use them to prove Theorem 1.

Recall $\overleftarrow{\mu}$ denotes the projection of μ to pasts, and let $\overleftarrow{\mu}_{x_\ell\ldots x_1}$ be the measure on pasts that precede the word

$w := x_1 \ldots x_\ell$. These are given by the word probabilities:

$$\begin{aligned} \Pr_{\overleftarrow{\mu}}(\, x_0 \ldots x_{-k}\,) &:= \Pr_\mu(\, x_{-k} \ldots x_0\,) \\ \Pr_{\overleftarrow{\mu}_{x_\ell \ldots x_1}}(\, x_0 \ldots x_{-k}\,) &:= \Pr_\mu(\, x_{-k} \ldots x_0 x_1 \ldots x_\ell\,) \ ; \end{aligned}$$

or, alternatively, by the measure values:

$$\begin{aligned} \overleftarrow{\mu}(U_{0,x_0 \ldots x_{-k}}) &:= \mu(U_{-k+1,x_{-k} \ldots x_0}) \\ \overleftarrow{\mu}_{x_\ell \ldots x_1}(U_{0,x_0 \ldots x_{-k}}) &:= \mu(U_{-k+1,x_{-k} \ldots x_0 x_1 \ldots x_\ell}) \ . \end{aligned}$$

Then Eq. (1) can be recast in the form of a convergence of likelihood ratios, taken over a sequence of cylinder sets $U_k := U_{0,x_0 \ldots x_{-k}}$ converging on $\overleftarrow{x}$:

$$\Pr_\mu\left(x_1 \ldots x_\ell \,\middle|\, \overleftarrow{X} \right) = \lim_{k \to \infty} \frac{\overleftarrow{\mu}_{x_\ell \ldots x_1}(U_k)}{\overleftarrow{\mu}(U_k)} \ . \qquad (4)$$

This reformulation, though somewhat conceptually cumbersome, is useful due to theorems that relate the convergence of likelihood ratios to the Radon-Nikodym derivative. Indeed, wherever Eq. (4) converges, it will be equal to the Radon-Nikodym derivative $d\overleftarrow{\mu}_{x_\ell \ldots x_1} / d\overleftarrow{\mu}(\overleftarrow{X})$.

To use these theorems we must define a *differentiation basis*. Any collection of neighborhoods $\mathcal{D}$ in $\mathcal{X}^{\mathbb{N}}$ may be considered a differentiation basis if for every $\overleftarrow{X} \in \mathcal{X}^{\mathbb{N}}$, there exists a sequence of neighborhoods (D_k) such that $\lim_{k \to \infty} D_k = \left\{ \overleftarrow{X} \right\}$. See Fig. 1.

The Vitali theorem states that whenever the differentiation basis $\mathcal{D}$ possesses the *Vitali property* with respect to "background" measure μ, then for μ-almost all $\overleftarrow{X}$ and any "test" measure ν which is absolutely continuous with respect to μ, the limit of likelihood ratios $\nu(V)/\mu(V)$ exists for any sequence $(V_k) \subset \mathcal{D}$ converging on $\overleftarrow{X}$ and is equal to the Radon-Nikodym derivative $d\nu/d\mu(\overleftarrow{X})$ at that point [31]. This sort of very flexible limit is denoted by:

$$\lim_{\substack{V \in \mathcal{D} \\ V \ni \overleftarrow{X}}} \frac{\nu(V)}{\mu(V)} = \frac{d\nu}{d\mu}(\overleftarrow{X}) \ .$$

The Vitali property has strong and weak forms, but we prove the strong form. Given a differentiation basis $\mathcal{D}$, a sub-differentiation basis $\mathcal{D}' \subseteq \mathcal{D}$ is any differentiation basis $\mathcal{D}'$ each of whose neighborhoods also belongs to $\mathcal{D}$. The differentiation basis $\mathcal{D}$ has the strong Vitali property with respect to μ if for every measurable set A and for every sub-differentiation basis $\mathcal{D}' \subseteq \mathcal{D}$ covering A, there is an *at most countable* subset $\{ D_j \} \subseteq \mathcal{D}'$ such that $D_j \cap D_{j'}$ is empty for all $j \neq j'$ and:

$$\mu\left(A - \left(\bigcup_j D_j \right) \right) = 0 \ .$$

In other words, we must be able to cover "almost all" of A with a countable number of nonoverlapping sets from the differentiation basis [31].

We now demonstrate that the differentiation basis $\mathcal{D}$ generated by cylinder sets on $\mathcal{X}^{\mathbb{N}}$ has the Vitali property for *any measure* μ.

Proposition 3 (Vitali property for stochastic processes)**.** *For any stochastic process $(\mathcal{X}^{\mathbb{N}}, \Sigma, \mu)$, let $\mathcal{D}$ be the differentiation basis of allowed cylinder sets. Then $\mathcal{D}$ has the strong Vitali property.*

Proof. *Let $\mathcal{D}' \subseteq \mathcal{D}$ be any subdifferentiation basis covering $\mathcal{X}^{\mathbb{N}}$. (Our proof trivially generalizes to any $A \subseteq \mathcal{X}^{\mathbb{N}}$.) Since $\mathcal{D}'$ is a differentiation basis, for all $\overleftarrow{X} \in \mathcal{X}^{\mathbb{N}}$ there must be a sequence $(D_j(\overleftarrow{X}))$ of cylinder sets converging on $\overleftarrow{X}$. Without loss of generality, we suppose that $D_j(\overleftarrow{X}) = U_{0,x_0 \ldots x_{-\ell_j+1}}$ with ℓ_j monotonically increasing. (If this is not the case, we take a subsequence of $D_j(\overleftarrow{X})$ for which it is the case.)*

Now consider the combination of all such sequences:

$$\mathcal{D}'' := \bigcup_{\overleftarrow{X} \in \mathcal{X}^{\mathbb{N}}} \left\{ D_j(\overleftarrow{X}) \,\middle|\, j \in \mathbb{N} \right\} \ .$$

We note that $\mathcal{D}''$, though a union of an uncountable number of sets, itself cannot be larger than a countable set, as the elements of the sets from which it is composed are characterized by finite words, and finite words themselves only form a countable set. That is, there is significant redundancy in $\mathcal{D}''$ that keeps it countable. Furthermore, $\mathcal{D}''$ has a lattice structure given by the set inclusion relation $\subseteq$ with the particular property that for $U, V \in \mathcal{D}''$, $U \cap V$ is nonempty only if $U \subseteq V$ or vice versa.

We then choose the set $\mathcal{C}$ of all maximal elements of this lattice: that is, those $U \in \mathcal{D}''$ such that there is no $V \in \mathcal{D}''$ containing U. These maximal elements must exist since for each $U \in \mathcal{D}''$ there is only a finite number of sets in $\mathcal{D}''$ that can contain it.

It must be the case that all sets in $\mathcal{C}$ are nonoverlapping. Furthermore, for any $V \in \mathcal{D}''$, not in $\mathcal{C}$, there can only be a finite number of such sets containing V. One of them must be maximal and therefore in $\mathcal{C}$. In particular, for every $\overleftarrow{X} \in \mathcal{X}^{\mathbb{N}}$, each of its neighborhoods in $\mathcal{D}''$ is contained by the union of $\mathcal{C}$.

This implies $\mathcal{C}$ is a complete covering of $\mathcal{X}^{\mathbb{N}}$. Since it is also nonoverlapping and countable, the strong Vitali property is proven.

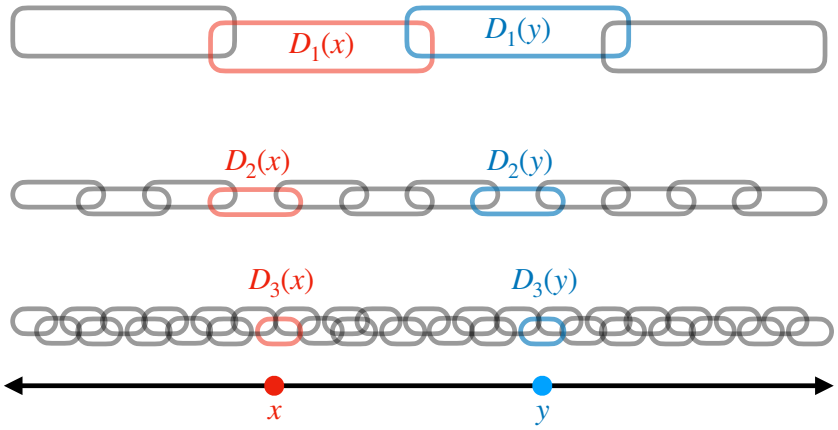

FIG. 1: *Snapshot of a differentiation basis:* A differentiation basis is a collection of neighborhoods in $\mathcal{X}^{\mathbb{N}}$ that have hierarchical structure. For every point $x \in \mathcal{X}^{\mathbb{N}}$, there must be a sequence of neighborhoods converging on that point. Pictured above, a line is shown with a partial representation of its differentiation basis above it in the form of a hierarchical collection of rounded rectangles. For two points x and y we show the corresponding sequence of sets $(D_j(x))$, $(D_j(y))$ converging on each.

As a consequence, the likelihood ratios in Eq. (4) must converge for $\overleftarrow{\mu}$-almost every past $\overleftarrow{X}$ and every finite-length word w—proving Theorem 1.

We note that this result follows as a relatively straightforward application of the Vitali property, which holds for any measure μ on $\mathcal{X}^{\mathbb{Z}}$ and $\mathcal{X}^{\mathbb{N}}$. Our good fortune is due to the particularly well-behaved topology of sequences of discrete observations. For continuous observations, a less direct path to predictive states must be taken.

B. Continuous observations: Overview

Shifting from discrete to real-valued observations, where now $\mathcal{X}$ denotes a compact subset of $\mathbb{R}^d$, multiple subtleties come to the fore.

First, it must be noted that even in $\mathbb{R}$, the existence of a Vitali property is not trivial. For the Lebesgue measure, only a weak Vitali property holds, though this is still sufficient for the equivalence between Radon-Nikodym derivatives and likelihood ratios. The differentiation basis in this setting can be taken to be comprised of all intervals (a, b) on the real line.

Second, to go from $\mathbb{R}$ to $\mathbb{R}^d$, constraints must be placed on the differentiation basis. An "interval" here is really the Cartesian product of intervals, but for a Vitali property to hold we must only consider products of intervals whose edges are held in a fixed ratio to one another, so that the edges converge uniformly to zero. Likelihood ratios for fixed-aspect boxes of this kind can converge to the Radon-Nikodym derivative [31].

This requirement poses a challenge for generalizing the Vitali property to infinite dimensions, as we must to study sequences of real numbers. A fixed-aspect "box" around a sequence of real numbers is not a practical construction. In the empirical setting, we can only observe information about a finite number of past outputs. We therefore cannot obtain any "uniform" knowledge of the entire past. That is, a direct generalization of the case for $\mathbb{R}^d$ does not suffice.

However, integration and differentiation on infinite-dimensional spaces has been considered before, mainly by Jessen [32, 33] and later Enomoto [34]. Their results focused on generalizing Lebesgue measure to $(S^1)^{\mathbb{N}}$, where S^1 is the circle. This section shows that their results can be significantly extended. The primary result we prove is a generalization of Enomoto's Theorem [34]:

Theorem 2 (Generalized Enomoto's Theorem)**.** *Let $\mathcal{X}$ be an interval of $\mathbb{R}$, and let μ be any probability measure over $\mathcal{X}^{\mathbb{N}}$. Let $f : \mathcal{X}^{\mathbb{N}} \to \mathbb{R}^+$ and let F be its indefinite integral under μ. Let $\mathcal{V}$ denote the differentiation basis consisting of sets of the form:*

$$V_{n,\delta}(\overleftarrow{X}) = \left\{ \overleftarrow{Y} \;\middle|\; |y_j - x_j| < \delta,\ j = 1, \ldots, n \right\} .$$

Then:

$$\lim_{\substack{V \in \mathcal{V} \\ V \ni \overleftarrow{X}}} \frac{F(V)}{\mu(V)} = f(\overleftarrow{X}) \ , \tag{5}$$

for $\overleftarrow{\mu}$-almost all $\overleftarrow{X}$.

Note that the resulting differentiation basis is a weaker form of that considered above. Each $V_{n,\delta}$ is evidently a cylinder set, but of a very particular kind. As we take $\delta \to 0$ and $n \to \infty$, we extend the "window" of the cylinder set to the entire past while simultaneously narrowing its width *uniformly*. This turns out to be sufficient to replicate the same effect as the fixed-aspect boxes in the finite-dimensional case.

As a corollary of Theorem 2, we have the following result for predictive states:

Corollary 2. *Let $\mathcal{X} \subset \mathbb{R}^d$ be a compact subset. For all measures μ on $\mathcal{X}^{\mathbb{Z}}$ and μ-almost every past $\overleftarrow{X} \in \mathcal{X}^{\mathbb{N}}$, define the measures $\eta_\ell[\overleftarrow{X}]$ and $\eta_{\ell,\delta}[\overleftarrow{X}]$ as:*

$$\eta_{\ell,\delta}[\overleftarrow{X}](U) = \frac{\mu(V_{\ell,\delta} \times U)}{\mu(V_{\ell,\delta})} \ \textit{and} \tag{6}$$

$$\eta_\ell[\overleftarrow{X}] = \lim_{\delta \to 0} \eta_{\ell,\delta}[\overleftarrow{X}] \ . \tag{7}$$

Then $\eta_\ell[\overleftarrow{X}] \to \epsilon[\overleftarrow{X}]$ in distribution, as $\ell \to \infty$.

Proof. *From Theorem 2 for any neighborhood U:*

$$\Pr_\mu \left(U \,\middle|\, \overleftarrow{X} \right) := \lim_{\ell \to \infty} \eta_{\ell,\delta(\ell)}[\overleftarrow{X}](U) \tag{8}$$

converges as long as $\delta(\ell) > 0$ for all ℓ and $\delta(\ell) \to 0$.

First, we must show that it is also perfectly fine to take the limit $\delta \to 0$ before taking $\ell \to \infty$. For each $\zeta > 0$, let $\Delta(\ell, \zeta)$ be chosen such that $\eta_{\ell,\Delta(\ell,\zeta)}[\overleftarrow{X}]$ is ζ-close to $\epsilon[\overleftarrow{x}]$, in the sense that:

$$\left| \Pr_{\eta_{\ell,\delta(\ell,\zeta)}[\overleftarrow{X}]}(U) - \Pr_{\eta_\ell[\overleftarrow{X}]}(U) \right| < \zeta \ .$$

Choose $\delta(\ell, \zeta) := \min\left\{ \Delta(\ell, \zeta), \ell^{-1} \right\}$ so that $\delta(\ell, \zeta) \to 0$. Then clearly $\eta_{\ell,\delta}[\overleftarrow{X}](U) \to \Pr_\mu \left(U \,\middle|\, \overleftarrow{X} \right)$, and this holds for all $\zeta > 0$. We therefore trivially have:

$$\Pr_\mu \left(U \,\middle|\, \overleftarrow{X} \right) = \lim_{\zeta \to 0} \lim_{\ell \to \infty} \Pr_{\eta_{\ell,\delta(\ell,\zeta)}[\overleftarrow{X}]}(U)$$

$$\eta_\ell[\overleftarrow{X}](U) = \lim_{\zeta \to 0} \Pr_{\eta_{\ell,\delta(\ell,\zeta)}[\overleftarrow{X}]}(U) \ .$$

And, this limit is in fact uniform by the construction of $\delta(\ell, \zeta)$. Therefore we may interchange the ζ and ℓ limits to get:

$$\begin{aligned} \Pr_\mu \left(U \,\middle|\, \overleftarrow{X} \right) &= \lim_{\ell \to \infty} \lim_{\zeta \to 0} \Pr_{\eta_{\ell,\delta(\ell,\zeta)}[\overleftarrow{X}]}(U) \\ &= \lim_{\ell \to \infty} \eta_\ell[\overleftarrow{X}](U) \end{aligned}$$

Last, we noted previously that this need only hold for a countable number of neighborhoods U in order for $\Pr_\mu \left(U \,\middle|\, \overleftarrow{X} \right)$ to generalize to a measure. We call this measure $\epsilon[\overleftarrow{X}]$ and, by Prop. 2, we have $\eta_\ell[\overleftarrow{X}] \to \epsilon[\overleftarrow{X}]$ in distribution for $\overleftarrow{\mu}$-almost all $\overleftarrow{X}$.

Note here that we allowed $\mathcal{X} \subset \mathbb{R}^d$. This can be obtained from Enomoto's theorem by simply reorganizing a sequence of d-dimensional coordinates from $(\mathbf{x}_1, \mathbf{x}_2, \dots)$ to $(x_{11}, \dots, x_{d1}, x_{12}, \dots, x_{d2}, \dots)$. Enomoto's theorem then requires uniformity of the intervals across past instances as well as within each copy of $\mathbb{R}^d$.

As before, the quantities $\Pr_\mu \left(U \,\middle|\, \overleftarrow{X} \right)$ define a unique measure $\epsilon[\overleftarrow{X}]$ on $\mathcal{X}^{\mathbb{N}}$. It is determined by:

$$\epsilon[\overleftarrow{X}](U) = \Pr_\mu \left(U \,\middle|\, \overleftarrow{X} \right) \ .$$

Enomoto's theorem itself is the capstone result in a sequence of theorems initiated by Jessen [32]. To prove Theorem 2, we must start from the beginning, generalizing Jessen's results. Fortunately, the bulk of the effort comes in generalizing the first of these results—Jessen's correspondence principle. After this, the generalization follows quite trivially from the subsequent theorems.

The next section provides the full proof for a generalized correspondence principle and explains how this result impacts the proofs of the subsequent theorems. For completeness, we also give the full proof of the generalized Enomoto's theorem, though it does not differ much from Enomoto's—published in French—once the preceding theorems are secured.

C. Jessen's correspondence principle

The Jessen and Enomoto theory rests on a profound correspondence between cylinder sets on $\mathcal{X}^{\mathbb{N}}$ and intervals on $\mathbb{R}$. To state it, we must define the concept of a net.

A net is similar to but formally separate from a differentiation basis, but like the latter allows for a notion of differentiation, called *differentiation-by-nets*. This is weaker than the Vitali property on a differentiation basis, but following on Jessen's work, Enomoto showed that differentiation-by-nets can be extended to describe a particular differentiation basis with the Vitali property.

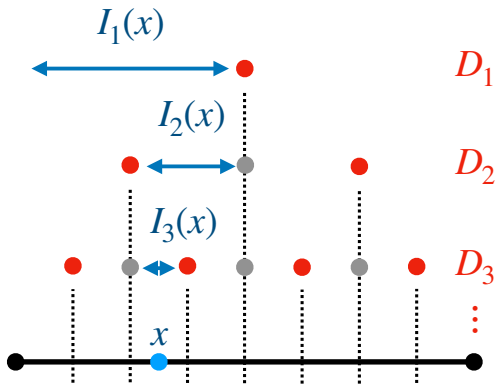

FIG. 2: *Snapshot of a differentiation net:* A differentiation net defined on a line segment. $D_1, D_2, D_3, \ldots$ represent the dissections that comprise the net. Each dissection contains the last. New points are indicated in red and old points in gray. These points define intervals; a sequence $(I_k(x))$ of these intervals is shown converging on the point x.

Let $\mathcal{X}$ be a finite interval on $\mathbb{R}$. A *dissection* $D = (b_1, \ldots, b_N)$ of $\mathcal{X}$ is simply a sequence of cut points, that generate a sequence of adjacent intervals (b_k, b_{k+1}) spanning $\mathcal{X}$, covering all but a finite set of points—the interval edges. See Fig. 2. Denote the intervals $\mathcal{I}(D) = \{ (b_k, b_{k+1}) \mid k = 1, \ldots, N-1 \}$. The length of the largest interval in $\mathcal{I}(D)$ is denoted $|D|$. (Not to be confused with D's cardinality, that we have no need to reference.) A *net* $\mathcal{N} = (D_n)$ is a sequence of dissections so that $D_n \subset D_{n+1}$ (that is, each new dissection only adds further cuts) and $|D_n| \to 0$ (the largest interval length goes to zero). The boundary $\partial\mathcal{N} = \bigcup_n D_n$ denotes all the boundary points from the sequence and is always a countable set.

We can similarly define a dissection $D = (d_1, \ldots, d_\ell)$ on $\mathcal{X}^{\mathbb{N}}$ as a set of ℓ dissections, one for each of the first ℓ copies of $\mathcal{X}$. D intervals $\mathcal{I}(D) = \left\{ i_1 \times \ldots i_\ell \times \mathcal{X}^{\mathbb{N}} \;\middle|\; i_k \in \mathcal{I}(d_k) \right\}$ are the cylinder sets generated by the intervals of each individual dissection. See Fig. 3. The boundary of a dissection is the set of all points that do not belong to these intervals: $\partial D = \left\{ X \in \mathcal{X}^{\mathbb{N}} \;\middle|\; \exists k : x_k \in d_k \right\}$. The size of the dissection is $|D| := \max_k |d_k|$. For a finite measure μ, there are always dissections with $\mu(\partial D) = 0$ of any given $|D| = \max_k |d_k|$, since $\mu|_{\mathcal{X}^\ell}$ can only have at most countably many singular points.

A net $\mathcal{N} = (D_n = (d_{1,n}, \ldots, d_{\ell_n,n}))$ of $\mathcal{X}^{\mathbb{N}}$ is a sequence of dissections of increasing depth ℓ_n so that each sequence $(d_{k,n})$ for fixed k is a net for the kth copy of $\mathcal{X}$. $\partial\mathcal{N} = \bigcup_n \partial D_n$ denotes all the accumulated boundary points of this sequence. Again, for finite measure μ, nets always exist that have $\mu(\partial\mathcal{N}) = 0$ for all n. Nets with this property are called *μ-continuous nets.*

Note that for any net, every sequence of intervals (I_n), $I_n \in \mathcal{I}(D_n)$ and $I_{n+1} \subset I_n$, uniquely determines a point $X \in \mathcal{X}^{\mathbb{N}}$. If $X \notin \partial D$, then X uniquely determines a sequence of intervals.

The following result can be proven—as a generalization from Ref. [32]:

Theorem 3 (Generalized correspondence principle)**.** *Let $\mathcal{X} \subset \mathbb{R}$ be an interval and let λ be the Lebesgue measure on $\mathcal{X}$, normalized so $\lambda(\mathcal{X}) = 1$. Let μ be a finite measure on $\mathcal{X}^{\mathbb{N}}$ that has no singular points. Let $\mathcal{N} = (D_n)$ be any μ-continuous net of $\mathcal{X}^{\mathbb{N}}$. Then there exists a net $\mathcal{M} = (d_n)$ of $\mathcal{X}$ so that:*

1. *There exists a function Φ_n that maps each interval in $\mathcal{I}(D_n)$ of positive measure to one and only one interval in $\mathcal{I}(d_n)$ and vice versa for Φ_n^{-1};*
2. *$\lambda(\Phi_n(I)) = \mu(I)$ for all $I \in \mathcal{I}(D_n)$ with $\mu(I) > 0$; and*
3. *The mapping $\phi : \mathcal{X}^{\mathbb{N}} - \partial\mathcal{N} \to \mathcal{X} - \partial\mathcal{M}$, generated by $X \mapsto (I_n) \mapsto (\Phi_n(I_n)) \mapsto x$, is measure-preserving.*

To summarize this technical statement: For any method of indefinitely dissecting the set $\mathcal{X}^{\mathbb{N}}$ into smaller and smaller intervals, there is in fact an "equivalent" such method for dissecting the much simpler set $\mathcal{X}$. It is equivalent in the sense that all the resulting intervals are in one-to-one correspondence with one another—a correspondence that preserves measure. Since interval sequences uniquely determine points (and vice versa for a set of full measure), this induces a one-to-one correspondence between points that is also measure-preserving.

The proof consists of two parts. The first proves the first two claims about $\mathcal{M}$. Namely, there is an interval correspondence and it is measure-preserving. The second shows this extends to a correspondence between $\mathcal{X}^{\mathbb{N}}$ and $\mathcal{X}$ that is also measure-preserving.

Proof (Interval correspondence)**.** *The proof proceeds by induction. For a given μ-continuous net $\mathcal{N} = (D_n)$, suppose we already constructed dissections $d_1, \ldots, d_N$ of $\mathcal{X}$ so that a function Φ_n between positive-measure intervals in D_n and d_n exists with the desired properties (1) and (2) above, for all $n = 1, \ldots, N$.*

Now, for D_{n+1}, a certain set of the intervals in $\mathcal{I}(D_n)$ is divided. Suppose $I \in \mathcal{I}_n$ divides into I' and I''. If

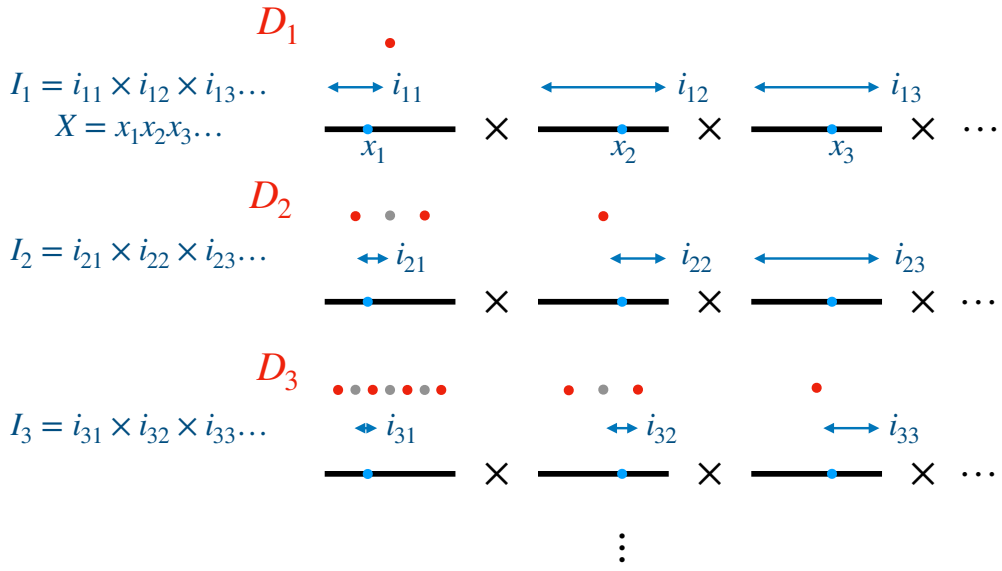

FIG. 3: *Snapshot of a differentiation net on a product space:* A differentiation net defined on a product space $\mathcal{X}^{\mathbb{N}}$. This is comprised of an increasingly detailed dissection on each factor space. Also shown is a sequence of product intervals converging on a point $X = x_1 x_2 x_3 \ldots$.

either of these, say I'', has measure zero then we discard it and set $\Phi_{n+1}(I') = \Phi_n(I)$. Otherwise, suppose that $\Phi_n(I) = (a, b)$. Then divide $\Phi_n(I)$ into the intervals:

$$\Phi_{n+1}(I') := \left(a, \frac{a\mu(I) + (b-a)\mu(I')}{\mu(I)}\right) \text{ and}$$
$$\Phi_{n+1}(I'') := \left(\frac{a\mu(I) + (b-a)\mu(I')}{\mu(I)}, b\right) ,$$

that clearly have Lebesgue measures $\lambda(\Phi_{n+1}(I')) = \mu(I')$ and $\lambda(\Phi_{n+1}(I'')) = \mu(I'')$, respectively. Generalizing this to more complicated divisions of I is straightforward.

We can always suppose for a given net $\mathcal{N}$ that D_0 is simply the trivial dissection that makes no cuts and only one interval. However, this has a trivial correspondence with $\mathcal{X}$; namely, $\Phi_0(\mathcal{X}^{\mathbb{N}}) = \mathcal{X}$.

By induction, then, the desired $\mathcal{M}$ can always be constructed.

With the existence of the interval correspondence established, we further demonstrate the existence of a point correspondence between μ-almost-all of $\mathcal{X}^{\mathbb{N}}$ and λ-almost-all of $\mathcal{X}$.

Proof (Point correspondence)**.** *For every $X \in \mathcal{X}^{\mathbb{N}} - \partial\mathcal{N}$, there is a unique sequence (I_n) of concentric intervals, $I_n \in \mathcal{I}(D_n)$ and $I_{n+1} \subset I_n$, such that $\bigcap_n I_n = \{X\}$. If X is in the support of μ, then we define:*

$$\phi(X) := \bigcap_n \Phi_n(I_n)$$

as the corresponding point in $\mathcal{X} - \partial\mathcal{M}$. Due to the interval correspondence, this mapping is invertible.

By measure-preserving we mean that for all $A \subseteq \mathcal{X}^{\mathbb{N}} - \partial\mathcal{N}$, $\lambda(\phi(A)) = \mu(A)$ and vice-versa for ϕ^{-1}. Both the Lebesgue measure and μ must be outer regular, due to being finite measures. Outer regular means that the measure of a set A is the infimum of the measure of all open sets containing A, a property we use to our advantage.

Consider for each n the minimal covering $\mathcal{C}_n$ of A by intervals in $\mathcal{I}(D_n)$. The measure of this covering is denoted $m_n := \mu(\bigcup \mathcal{C}_n)$. Clearly, $m_n \geq \mu(A)$ and $m_n \to \mu(A)$. The corresponding covering $\Phi_n(\mathcal{C}_n)$ in $\mathcal{I}(d_n)$ is a covering of $\phi(A)$ and has the same measure m_n. By outer regularity, then, $m_n \geq \lambda(\phi(A))$ for all n. And so, $\mu(A) \geq \lambda(\phi(A))$.

Now, by the exact reverse argument of the previous paragraph, going from $\mathcal{X}$ to $\mathcal{X}^{\mathbb{N}}$ via ϕ^{-1}, we can also deduce that $\mu(A) \leq \lambda(\phi(A))$. Therefore $\mu(A) = \lambda(\phi(A))$, and the function ϕ is measure-preserving.

D. Corollaries and Enomoto's Theorem

Jessen's correspondence principle is an extremely powerful device. Among its consequences are the following theorems regarding functions on $\mathcal{X}^{\mathbb{N}}$. We state their generalized forms here and for the proofs refer to Jessen [32], as each is a direct application of Theorem 3 without making any further assumptions on the measure μ.

The first offers a much weaker (and on its own, insufficient for our purposes) concept of differentiation of measures that we refer to as *differentiation-by-nets*.

Corollary 3 (Differentiation-by-nets)**.** *Let $f : \mathcal{X}^{\mathbb{N}} \to \mathbb{R}^+$ and let F be the measure defined by its indefinite integral: $F(A) := \int_A f(X)d\mu(X)$. Further let $\mathcal{N} = (D_n)$ be a net on $\mathcal{X}^{\mathbb{N}}$ and denote by $\hat{f}_n$ a piecewise function such that $\hat{f}_n(X) = F(I_n)/\mu(I_n)$ for all $X \in I_n$ and each $I_n \in D_n$. Then $\hat{f}_n(X) \to f(X)$ as $n \to \infty$ for μ-almost all X.*

Though the full proof is found in Ref. [32], we summarize its key point: Using the correspondence of intervals, we write $F(I_n)/\mu(I_n) = \tilde{F}(\Phi(I_n))/\lambda(\Phi(I_n))$, where $\tilde{F}$ is the indefinite integral of $f \circ \phi^{-1}$ with respect to λ. The limit then holds due to the Vitali property of λ on $\mathcal{X}$. However, we also note that Corollary 3 is *not* an extension of the Vitali property to cylinder sets on $\mathcal{X}^{\mathbb{N}}$. Jessen himself offers a counterexample to this effect in a later publication [33].

Jessen's second corollary is key to demonstrating that $\mathcal{V}$, the differentiation basis defined in Theorem 2, *will* have the sought-after Vitali property.

Corollary 4 (Functions as limits of integrals)**.** *Let $f : \mathcal{X}^{\mathbb{N}} \to \mathbb{R}^+$, and let $f_n(X)$ be a sequence of functions given by:*

$$f_n(x_1x_2\dots) := \int_{Y\in\mathcal{X}^{\mathbb{N}}} f(x_1\dots x_nY)d\mu(Y) \ .$$

That is, we integrated over all observations after the first n. Thus, f_n only depends on the first n observations. Then $f_n(X) \to f(X)$ as $n \to \infty$ for μ-almost all X.

This proof we also skip, again referring the reader to Jessen [32], as no step is directly dependent on the measure μ itself and only on properties already proven by the previous theorems.

We now have sufficient knowledge to prove the generalized Enomoto's theorem; generalized from Ref. [34].

Proof (Generalized Enomoto's Theorem)**.** *First, we must demonstrate, for almost every X, that there exists a sequence $V_j(X)$ converging on X such that the limit holds. By Corollary 4, there must be, for μ-almost all X and any $\epsilon > 0$, a $k(X,\epsilon)$ such that $|f_n(X) - f(X)| < \epsilon/2$ for all $n > k(X,\epsilon)$. Now, from the Vitali property on μ_n and the fact that f_n only depends on the first n observations, it must be true that for any $\epsilon > 0$ and almost all X, there is a $0 < \Delta(X,n,\epsilon) < 1$ so that:*

$$\left| f_n(X) - \frac{F(V_{n,\delta}(X))}{\mu(V_{n,\delta}(X))} \right| < \epsilon/2 \ ,$$

whenever $\delta < \Delta(X,n,\epsilon)$. For a given ϵ, there is a countable number of conditions (one for each n). As such, the set of points X for which all conditions hold is still measure one. Then, taking for each X the integer $K := k(X,\epsilon)$ and subsequently the number $\Delta := \Delta(X,k(X,\epsilon),\epsilon)$, we can choose $V_{K,\Delta}(X)$ and by the triangle inequality we must have:

$$\left| f(X) - \frac{F(V_{K,\Delta}(X))}{\mu(V_{K,\Delta}(X))} \right| < \epsilon \ . \tag{9}$$

This completes the proof's first part.

However, the second part—that all *sequences $V_{n_j,\delta_j}(x)$ of neighborhoods give converging likelihood ratios—further follows from the above statements, as:*

$$\left| f(X) - \frac{F(V_{n_j,\delta_j}(X))}{\mu(V_{n_j,\delta_j}(X))} \right| < \epsilon$$

must hold for any $n_j > K(X,\epsilon)$ and any $\delta_j < \Delta(X,K(X,\epsilon),\epsilon)$, which must eventually be true for any converging sequence to X.

Now, the previous theorem does not directly prove the Vitali property but rather bypasses it. Demonstrating that the differentiation basis $\mathcal{V}$ may be used to recover Radon-Nikodym derivatives. This, then, is sufficient for Corollary 2 to hold, guaranteeing the existence of predictive states $\epsilon(\overleftarrow{X})$ for μ-almost all $\overleftarrow{X}$.

E. Convergence of predictive states

An important task regarding predictive states is to learn a process's predictive states—that is, the ϵ-mapping from observed pasts to distributions over futures—from a sufficiently large sample of observations. These learned predictive states may then be used to more accurately predict the process' future behavior based on behaviors already observed.

The previous three sections avoided constraining the measure μ on $\mathcal{X}^{\mathbb{N}}$ to be anything other than finite. It was not assumed to be either stationary or ergodic, in particular. In such cases the ϵ-mapping is time-dependent, as it obviously depends on where futures are split from

pasts. To adequately reconstruct $\epsilon[\overleftarrow{X}]$ from a single, long observation requires that the process be both stationary and ergodic. Stationarity makes $\epsilon[\overleftarrow{X}]$ time-independent, and ergodicity ensures that the probabilities in the limits Eqs. (1) and (8) can be approximated by taking the time-averaged frequencies of occurrence.

The next natural question is how rapidly convergence occurs for each past, in a given process. So far, we only guaranteed that convergence exists, but said nothing on its rate. This is process-dependent. Section V gives several examples of processes and process types with their convergence rate. The most useful way to think of the rate is in the form of "probably-almost-correct"-type statements, as exemplified in the following result:

Proposition 4. *Let μ be a probability measure on $\mathcal{X}^{\mathbb{Z}}$. Let $\eta_\ell[\overleftarrow{X}](U)$ be defined as in either Cors. 1 or 2. For every cylinder set U and $\Delta_1, \Delta_2 > 0$, we have for sufficiently large ℓ:*

$$\Pr_{\overleftarrow{\mu}}\left(\left|\eta_\ell[\overleftarrow{X}](U) - \epsilon[\overleftarrow{X}](U)\right| > \Delta_1\right) < \Delta_2 \ .$$

That is, the probability of an error beyond Δ_1 is less than Δ_2.

This is a consequence of the fact that all $\overleftarrow{X}$ must eventually converge. The possible relationships between Δ_1, Δ_2, and ℓ in particular is explored in our examples.

IV. PREDICTIVE STATES FORM A HILBERT SPACE

Thus far, we demonstrated that for discrete and real $\mathcal{X}$, measures over $\mathcal{X}^{\mathbb{Z}}$ possess a well-defined feature called *predictive states* that relate how past observations constrain future possibilities. These states are defined by convergent limits that can be approximated from empirical time series in the case of stationary, ergodic processes.

We turn our attention now to the topological and geometric structure of these states, the spaces they live in, and how the structure of these spaces may be leveraged in the inference process. The results make contact between predictive states *as elements of a Hilbert space* and the well-developed arena of reproducing kernel Hilbert spaces. To do this we introduce several new concepts.

Denote the set of real-valued continuous functions on $\mathcal{X}^{\mathbb{Z}}$ by $C(\mathcal{X}^{\mathbb{Z}})$. The set of signed measures on $\mathcal{X}^{\mathbb{Z}}$, that we call $\mathbb{M}(\mathcal{X}^{\mathbb{Z}})$, may be thought of as dual to $C(\mathcal{X}^{\mathbb{Z}})$. This allows us to define a notion of convergence of measures on $\mathcal{X}^{\mathbb{Z}}$ in relation to continuous functions. We say that a sequence of measures μ_n *converges in distribution* if:

$$\lim_{n\to\infty} \int F(X) d\mu_n(X) = \int F(X) d\mu(X)$$

for all $F \in C(\mathcal{X}^{\mathbb{Z}})$. Convergence in distribution is sometimes referred to as weak convergence but we avoid this vocabulary to minimize confusion—as another, distinct kind of weak convergence is needed in the Hilbert space setting.

A kernel $k : \mathcal{X}^{\mathbb{Z}} \times \mathcal{X}^{\mathbb{Z}} \to \mathbb{R}$ generates a *reproducing kernel Hilbert space* (RKHS) $\mathcal{H}$ if $k(\cdot,\cdot)$ is positive semi-definite and symmetric [35]. $\mathcal{H}$ is typically defined as a space of *functions* (from $\mathcal{X}^{\mathbb{Z}} \to \mathbb{R}$), but the kernel allows embedding measures on $\mathcal{X}^{\mathbb{Z}}$ into the function space through $f_\mu(x) = \int k(x,y) d\mu(y)$. This elicits an inner product between any two positive measures μ and ν:

$$\langle f_\mu | f_\nu \rangle_k := \int\int k(x,y) d\mu(x) d\nu(y) \ .$$

The inner-product space on measures generated by this construction is isometric to the RKHS generated by $k(\cdot,\cdot)$. The embedding of measures into this space is unique if the kernel is *characteristic*. And, convergence in the norm of the Hilbert space is equivalent to convergence in distribution whenever the kernel is *universal* [36].

What exactly is the set $\mathcal{H}$ of functions? The equivalence of convergence in norm and convergence in distribution tempts identifying $\mathcal{H}$ with the space of continuous functions, but this is overly optimistic. If it were true—that $\mathcal{H} = C(\mathcal{X}^{\mathbb{Z}})$—then the convergence $\langle F | f_{\mu_n} \rangle_k \to \langle F | f_\mu \rangle_k$ for every $F \in \mathcal{H}$ implies $f_{\mu_n} \to f_\mu$ in the norm topology of $\mathcal{H}$. That, though, would identify norm convergence on the Hilbert space with *weak* convergence, which for Hilbert spaces is identified as the convergence of every inner product. For infinite-dimensional Hilbert spaces, these two types of convergence cannot be identified, as is seen in the simple case of any orthonormal basis e_i, for which $\langle F | e_i \rangle \to 0$ is necessary for F to have a finite norm, even though $\|e_i\| \to 1$ by definition. So, we must conclude that while convergence in the norm of $\mathcal{H}$ is equivalent to the convergence in distribution of measures, $\mathcal{H}$ can only be a proper subspace of the continuous functions (a fact also noted in [37]). Weak convergence in $\mathcal{H}$ is then indeed weaker than convergence in distribution. (Hence, we do not call the latter "weak".)

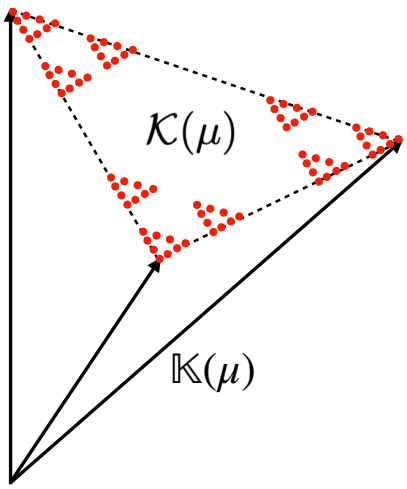

FIG. 4: *Predictive states and their closed span.* The red dots are a hypothetical set $\mathcal{K}(\mu)$ of predictive states (shaped like the Sierpinski set) that is uncountably infinite but that has a finite-dimensional closed span $\mathbb{K}(\mu)$.

A. Topology of predictive states

Let the (closure of the) set of a process's predictive states be denoted by:

$$\mathcal{K}(\mu) := \overline{\left\{ \epsilon(\overleftarrow{X}) \;\middle|\; \overleftarrow{X} \in \mathcal{X}^{\mathbb{N}} \right\}} .$$

The closure is taken under convergence in distribution.

The relation between pasts and predictive states may be highly redundant. For instance, in the process generated by the results of a random coin-toss, since the future observations do not depend on past observations, $\mathcal{K}(\mu)$ is trivial. Meanwhile, for a periodic process of period k, $\mathcal{K}(\mu)$ has k elements, corresponding to the k distinct states—the process' phases. In more complex cases, $\mathcal{K}(\mu)$ may have countable and uncountable cardinality.

We may also consider the vector space of signed measures generated by the closed span of $\mathcal{K}(\mu)$, denoted $\mathbb{K}(\mu)$. This is the smallest closed vector space that contains the predictive states. This, too, may demonstrate redundancy in the form of linear dependence, regardless of the cardinality of $\mathcal{K}(\mu)$. For instance, it is possible to have an uncountably infinite set of predictive states $\mathcal{K}(\mu)$ whose dimension, $\dim \mathbb{K}(\mu)$, is finite. In fact, it is the general case that any process generated by an HMM or GHMM will have finite dimensional $\mathbb{K}(\mu)$, but is not guaranteed to have finite $\mathcal{K}(\mu)$ [12]. See Fig. 4.

The topology of convergence in distribution is closely related to the definition of continuity on $\mathcal{X}^{\mathbb{N}}$. It behooves us at this juncture to discuss $\mathcal{X}^{\mathbb{N}}$ not only as a topological space but also as a metric space.

Two useful families of distance metrics, equivalent to the product topology on $\mathcal{X}^{\mathbb{N}}$, are the Euclidean metrics, one for the discrete and real case each:

$$D_{\mathrm{E},\gamma}(X,Y)^2 := \begin{cases} \sum_{t=1}^{\infty} (1 - \delta_{x_t y_t}) \gamma^{2t} & \mathcal{X} \text{ discrete} \\ \sum_{t=1}^{\infty} \|x_t - y_t\|^2 \gamma^{2t} & \mathcal{X} \subset \mathbb{R}^d \end{cases} ,$$

for some $0 < \gamma < 1$. These distance metrics arise from embedding $\mathcal{X}^{\mathbb{N}}$ in a Hilbert space. Given an orthogonal basis (e_i), the components of this embedding for the discrete case are given by:

$$c_i(X) = \begin{cases} \gamma^{\lfloor i/|\mathcal{X}| \rfloor} & x_{\lfloor i/|\mathcal{X}| \rfloor} = i \bmod |\mathcal{X}| \\ 0 & \text{otherwise} \end{cases}$$

and in the continuous case ($\mathcal{X} \subset \mathbb{R}^d$) by:

$$c_i(X) = \gamma^t x_{k,t}, \quad i = k \mod d .$$

Using these distance metrics, the following section introduces an inner product structure on $\mathbb{K}(\mu)$ whose norm metrizes the topology of convergence in distribution. Once $\mathbb{K}(\mu)$'s natural embedding into a Hilbert space of its own is established, we investigate how well this embedding can be approximated by the approach of reproducing kernel Hilbert spaces.

We note that the $D_{\mathrm{E},\gamma}$ family of metrics is not the only family that metrizes the product topology on $\mathcal{X}^{\mathbb{N}}$. Further, though each $D_{\mathrm{E},\gamma}$ for $0 < \gamma < 1$ metrizes the product topology, they are not metrically equivalent in the strong sense. Thus, while the precise choice of metric is irrelevant for our present scope—to demonstrate the topological appropriateness of RKHS algorithms—further investigation into their metric properties is warranted. We discuss these issues in Supplementary Material II.

B. Embedding predictions in a Hilbert space

The space $\mathbb{K}(\mu)$ of predictive states is a subspace $\mathbb{P}(\mathcal{X}^{\mathbb{N}})$ of the probability measures over $\mathcal{X}^{\mathbb{N}}$. On $\mathbb{P}(\mathcal{X}^{\mathbb{N}})$, given any symmetric positive-definite kernel $k : \mathcal{X}^{\mathbb{N}} \times \mathcal{X}^{\mathbb{N}} \to \mathbb{R}$, we can define an inner product over measures:

$$\langle \mu | \nu \rangle_k := \int \int k(X,Y) d\mu(X) d\nu(Y) . \qquad (10)$$

Positive-definite means that for any finite set $\{X_i\}$ of $X_i \in \mathcal{X}^{\mathbb{N}}$ and any set $\{c_i\}$ of values $c_i \in \mathbb{R}$, both sets

have the same cardinality:

$$\sum_{i,j} k(X_i, X_j) c_i c_j \geq 0 \ ,$$

with equality only when $c_i = 0$ for all i. If this is true, then the inner product Eq. (10) is positive-definite for all *measures*. That is, $\langle \mu | \mu \rangle_k \geq 0$ with equality only when $\mu = 0$ [36].

Since $\mathcal{X}^{\mathbb{N}}$ is compact, if the kernel k satisfies the property of being *universal*, then norm convergence under the inner product defined by k is equivalent to convergence in distribution of measures [36].

A simple example of a universal kernel is the Gaussian radial basis function, when paired with an appropriate distance—namely, one defined from embedding $\mathcal{X}^{\mathbb{N}}$ in a Hilbert space, as our $D_{\mathrm{E},\gamma}$ are [38]. These take the form:

$$k_{\beta,\gamma}(X, Y) := \exp\left(-\frac{D_{\mathrm{E},\gamma}(X, Y)^2}{\beta^2}\right) \ .$$

We denote the associated inner products by $\langle \cdot | \cdot \rangle_{\beta,\gamma}$. $\mathcal{H}_{\beta,\gamma} := \left(\mathbb{P}(\mathcal{X}^{\mathbb{N}}), \langle \cdot | \cdot \rangle_{\beta,\gamma}\right)$ defines a Hilbert space, since it has the topology of convergence in distribution and $\mathbb{P}(\mathcal{X}^{\mathbb{N}})$ is complete in this topology.

When referring to a measure μ as an element of $\mathcal{H}_{\beta,\gamma}$ we denote it $|\mu\rangle_{\beta,\gamma}$ and inner products in the bra-ket are $\langle \mu | \nu \rangle_{\beta,\gamma}$. Now, it should be noted that to every ket $|\mu\rangle_{\beta,\gamma}$ there is a bra $\langle \mu |_{\beta,\gamma}$ that denotes a dual element. However, the dual elements of $\mathbb{P}(\mathcal{X}^{\mathbb{N}})$ correspond to continuous functions. The function f_μ corresponding to $\langle \mu |_{\beta,\gamma}$ is given by:

$$f_\mu(Y) := \int k_{\beta,\gamma}(X, Y) d\mu(X) \ , \tag{11}$$

so that:

$$\langle \mu | \nu \rangle_{\beta,\gamma} = \int f_\mu(X) d\nu(Y) \ .$$

Let $\mathcal{F}_{\beta,\gamma}$ denote the space of all f_μ that can be constructed from Eq. (11). This function space, when paired with the inner product $\langle f_\mu | f_\nu \rangle^{\mathcal{F}}_{\beta,\gamma} := \langle \nu | \mu \rangle_{\beta,\gamma}$, is isomorphic to $\mathcal{H}_{\beta,\gamma}$. $\mathcal{F}_{\beta,\gamma}$ is then a reproducing kernel Hilbert space with kernel $k_{\beta,\gamma}$.

As the start of Section IV discussed, $\mathcal{F}_{\beta,\gamma} \subset C(\mathcal{X}^{\mathbb{N}})$. Furthermore, the $\mathcal{F}_{\beta,\gamma}$ are not identical to one another, obeying the relationship $\mathcal{F}_{\beta,\gamma} \subset \mathcal{F}_{\beta',\gamma}$ when $\beta > \beta'$ [39]. However, it is also the case that each $\mathcal{F}_{\beta,\gamma}$ is dense in $C(\mathcal{X}^{\mathbb{N}})$, so their representative capacity is still quite strong [36].

We note an important rule regarding the scaling of our inner products, as constructed. The distances $D_{\mathrm{E},\gamma}(X, Y)$ have finite diameter on our spaces. Let Δ denote the diameter of $\mathcal{X}$. For discrete $\mathcal{X}$ we simply have $\Delta = 1$; for $\mathcal{X} \subset \mathbb{R}^d$ it is determined by the Euclidean distance. Then $\mathcal{X}^{\mathbb{N}}$'s diameter is given by $\Delta / \sqrt{1 - \gamma^2}$. Let $u := \mu - \nu$. Using a Taylor expansion for the Gaussian, for arbitrarily large β:

$$\begin{aligned}
&\|\mu - \nu\|^2_{\beta,\gamma} \\
&\quad = \int \int k_{\beta,\gamma}(X, Y) du(X) du(Y) \\
&\quad \leq \int \int \left(1 - \frac{D_{\mathrm{E},\gamma}(X, Y)^2}{\beta^2} + O(\beta^{-3})\right) du(X) du(Y) \\
&\quad \leq \frac{2 \|\mu - \nu\|_{\mathrm{TV}} \Delta^2}{(1 - \gamma^2)\beta^2} + O(\beta^{-3}) \ ,
\end{aligned} \tag{12}$$

where $\| \cdot \|_{\beta,\gamma}$ is simply the norm of $\mathcal{H}_{\beta,\gamma}$ and $\| \cdot \|_{\mathrm{TV}}$ is the total variation norm. The first inequality follows from applying the Taylor expansion and the second from using the bounded magnitude of the $O(\beta^{-2})$ term (due to the diameter) in conjunction with the total variation of $u \otimes u$, which satisfies $\|u \otimes u\|_{\mathrm{TV}} \leq 2 \|u\|_{\mathrm{TV}}$. (The constant term vanishes under integration over the difference measure u.)

This tells us that the norm is less discriminating between measures as $\beta \to \infty$. Naturally, this can be remedied by rescaling the kernel with a β^2 factor. As it happens, Eq. (12) will be useful later.

C. Finite-length embeddings

Our goal is to study how reproducing kernel Hilbert spaces may be used to encode information about predictive states gleaned from empirical observations. Given that such observations are always finite in length, we must determine whether and in what manner the Hilbert space representations of measures over finite-length observations converges to the Hilbert space representation of a measure over infinite sequences.

Let $\mu_{|\ell}$ denote the measure μ restricted to $\mathcal{X}^\ell$. Define the restricted distance on $\mathcal{X}^\ell$:

$$D^{(\ell)}_{\mathrm{E},\gamma}(X, Y)^2 := \begin{cases} \sum_{t=1}^{\ell} (1 - \delta_{x_t y_t}) \gamma^{2t} & \mathcal{X} \text{ discrete} \\ \sum_{t=1}^{\ell} \|x_t - y_t\|^2 \gamma^{2t} & \mathcal{X} \subset \mathbb{R}^d \end{cases} \ ,$$

for $X, Y \in \mathcal{X}^\ell$. This gives an important Pythagorean theorem for sequences:

$$\begin{aligned}
D_{\mathrm{E},\gamma}(X, Y)^2 = &D^{(\ell)}_{\mathrm{E},\gamma}(x_1 \ldots x_\ell, y_1 \ldots y_\ell)^2 \\
&+ \gamma^{2\ell} D_{\mathrm{E},\gamma}(x_{\ell+1} \ldots, y_{\ell+1} \ldots)^2 \ .
\end{aligned} \tag{13}$$

Now, using $D^{(\ell)}_{\mathrm{E},\gamma}$ define kernels $k^{(\ell)}_{\beta,\gamma}$ in the same style as for $\mathcal{X}^{\mathbb{N}}$. These generate inner products on $\mathbb{P}(\mathcal{X}^\ell)$. Denote by $\mathcal{H}^{(\ell)}_{\beta,\gamma}$ the resulting Hilbert spaces. These are related to the original $\mathcal{H}_{\beta,\gamma}$ by the following factorization theorem:

Proposition 5. *The predictive Hilbert space $\mathcal{H}_{\beta,\gamma}$ factors into $\mathcal{H}^{(\ell)}_{\beta,\gamma} \otimes \mathcal{H}_{\beta\gamma^{-\ell},\gamma}$.*

Before stating the proof, we should explain the above. The factorization $\mathcal{H}_{\beta,\gamma} = \mathcal{H}^{(\ell)}_{\beta,\gamma} \otimes \mathcal{H}_{\beta\gamma^{-\ell},\gamma}$ denotes separating the infinite-dimensional $\mathcal{H}_{\beta,\gamma}$ into two pieces—one of which is finite-dimensional, but retains the same kernel parameters, and another piece that *reparametrizes* infinite-dimensional Hilbert space. The reparametrization is $\beta \to \beta\gamma^{-\ell}$. This constitutes, essentially, a renormalization-group technique, in which the topology of words starting at depth ℓ is equivalent to a reparametrization of the usual topology. This reparametrization works precisely due to the Pythagorean theorem for sequences Eq. (13).

Proof. *We are demonstrating an isomorphism—a particularly natural one. Let δ_X be the Dirac delta measure concentrated on X. We note that for any measure μ:*

$$|\mu\rangle_{\beta,\gamma} = \int |\delta_X\rangle_{\beta,\gamma} \, d\mu(X) \ .$$

Now, consider the linear function from $\mathcal{H}_{\beta,\gamma}$ to $\mathcal{H}^{(\ell)}_{\beta,\gamma} \otimes \mathcal{H}_{\beta\gamma^{-\ell},\gamma}$ that maps:

$$|\delta_X\rangle_{\beta,\gamma} \mapsto |\delta_{x_1 \ldots x_\ell}\rangle^{(\ell)}_{\beta,\gamma} \otimes |\delta_{x_{\ell+1}\ldots}\rangle_{\beta\gamma^{-\ell},\gamma} \ , \tag{14}$$

for every X. Then by Eq. (13) *we can see that:*

$$\begin{aligned}
&\langle \delta_{y_1 \ldots y_\ell} | \delta_{x_1 \ldots x_\ell} \rangle^{(\ell)}_{\beta,\gamma} \langle \delta_{y_{\ell+1}\ldots} | \delta_{x_{\ell+1}\ldots} \rangle_{\beta\gamma^{-\ell},\gamma} \\
&= e^{-\beta^{-2} D^{(\ell)}_{\mathrm{E},\gamma}(x_1 \ldots x_\ell, y_1 \ldots y_\ell)^2} e^{-\beta^{-2}\gamma^{2\ell} D_{\mathrm{E},\gamma}(x_{\ell+1}\ldots, y_{\ell+1}\ldots)^2} \\
&= e^{-\beta^{-2} D_{\mathrm{E},\gamma}(X,Y)} = \langle \delta_Y | \delta_X \rangle_{\beta,\gamma} \ ,
\end{aligned}$$

so the mapping Eq. (14) *preserves the inner product and thus is an isomorphism.*

Note that for any of these Hilbert spaces there exists an element corresponding to the constant function $\mathbf{1}(X) = 1$ for all X. This function always exists in $\mathcal{F}_{\beta,\gamma}$. We denote its corresponding measure in $\mathcal{H}_{\beta,\gamma}$ as $\lambda_{\beta,\gamma}$, so that $\langle \lambda_{\beta,\gamma} | \mu \rangle_{\beta,\gamma} = 1$ for all μ. Then the operator $\Pi^{(\ell)}_{\beta,\gamma} : \mathcal{H}_{\beta,\gamma} \to \mathcal{H}^{(\ell)}_{\beta,\gamma}$ is given by:

$$\Pi^{(\ell)}_{\beta,\gamma} := I^{(\ell)} \otimes \langle \lambda_{\beta,\gamma} |_{\beta,\gamma} \ ,$$

where $I^{(\ell)}$ is the identity on $\mathcal{H}^{(\ell)}_{\beta,\gamma}$. It provides the canonical mapping from a measure μ to its projection $\mu_{|\ell}$: That is, $\Pi^{(\ell)}_{\beta,\gamma} |\mu\rangle_{\beta,\gamma} = |\mu_{|\ell}\rangle^{(\ell)}_{\beta,\gamma}$.
Consider the "truncation error"—that is, the residual error remaining when representing a measure by its truncated form $\mu_{|\ell}$ rather than by its full form μ. We quantify this in terms of an embedding. That is, there exists an embedding of truncated measures $\mathbb{P}(\mathcal{X}^\ell)$ into the space of full measures $\mathbb{P}(\mathcal{X}^{\mathbb{N}})$ such that the distance between any full measure and its truncated embedding is small:

Theorem 4. *There exist isometric embeddings $\mathcal{H}^{(\ell)}_{\beta,\gamma} \mapsto \mathcal{H}^{(\ell')}_{\beta,\gamma}$ and $\mathcal{H}^{(\ell)}_{\beta,\gamma} \mapsto \mathcal{H}_{\beta,\gamma}$ for any $\ell \leq \ell'$. Furthermore, let μ be any measure and $\mu_{|\ell}$ be its projection to the first ℓ observations, and let $|\hat{\mu}_\ell\rangle_{\beta,\gamma}$ be the embedding of $\mu_{|\ell}$ into $\mathcal{H}_{\beta,\gamma}$. Then $|\hat{\mu}_\ell\rangle_{\beta,\gamma} \to |\mu\rangle_{\beta,\gamma}$ as $\ell \to \infty$, with $\|\mu - \hat{\mu}_\ell\|_{\beta,\gamma} \sim O(\beta^{-1}\gamma^\ell)$.*

Proof. *For a measure μ with projection $\mu_{|\ell}$ let $\hat{\mu}_\ell$ denote the measure on $\mathcal{X}^{\mathbb{N}}$ with the property:*

$$\hat{\mu}_\ell(A \times B) = \mu_{|\ell}(A) \lambda_{\beta\gamma^{-\ell},\gamma}(B) \ ,$$

for $A \in \mathcal{X}^\ell$ and $B \in \mathcal{X}^{\mathbb{N}}$. Then the mapping $\mu_{|\ell} \mapsto \hat{\mu}_\ell$ is simply a rescaling, since:

$$\begin{aligned}
\langle \hat{\mu}_\ell | \hat{\nu}_\ell \rangle_{\beta,\gamma} &= \int\int k_{\beta,\gamma}(X,Y) d\hat{\mu}_\ell(X) d\hat{\nu}_\ell(X) \\
&= \int\int k^{(\ell)}_{\beta,\gamma}(x_1 \ldots x_\ell, y_1 \ldots y_\ell) d\mu_\ell d\nu_\ell \\
&\quad \times \int\int k_{\beta\gamma^{-\ell},\gamma}(x_\ell \ldots, y_\ell \ldots) d\lambda_{\beta\gamma^{-\ell},\gamma} d\lambda_{\beta\gamma^{-\ell},\gamma} \\
&= \langle \mu_{|\ell} | \nu_{|\ell} \rangle^{(\ell)}_{\beta,\gamma} \int d\lambda_{\beta\gamma^{-\ell},\gamma} \\
&= \langle \mu_{|\ell} | \nu_{|\ell} \rangle^{(\ell)}_{\beta,\gamma} \left\| \lambda_{\beta\gamma^{-\ell},\gamma} \right\|^2_{\beta\gamma^{-\ell},\gamma} \ .
\end{aligned}$$

Now, as a result of Eq. (13)*, note that for any two measures μ and ν:*

$$\langle \mu | \nu \rangle_{\beta,\gamma} = \int d\mu_{|\ell}(x_1 \ldots x_\ell) \int d\nu_{|\ell}(y_1 \ldots y_\ell) \exp\left(-\beta^2 D^{(\ell)}_\gamma (x_1 \ldots x_\ell, y_1 \ldots y_\ell) \right) \langle \mu(\cdot | x_1 \ldots x_\ell), \nu(\cdot | y_1 \ldots y_\ell) \rangle_{\beta\gamma^\ell,\gamma}$$

If we combine this fact with the bound Eq. (12)*, we have the result:*

$$\begin{aligned}\left\|\mu - \hat{\mu_{|\ell}}\right\|^2_{\beta,\gamma} &= \int d\mu_{|\ell}(x_1 \ldots x_\ell) \int d\mu_{|\ell}(y_1 \ldots y_\ell) \\ &\quad \exp\left(-\beta^2 D^{(\ell)}_\gamma(x_1 \ldots x_\ell, y_1 \ldots y_\ell)\right) \left\|\mu(\cdot|x_1 \ldots x_\ell) - \hat{\mu}_\ell(\cdot|y_1 \ldots y_\ell)\right\|^2_{\beta\gamma^{-\ell},\gamma} \\ &\leq \int d\mu_{|\ell}(x_1 \ldots x_\ell) \int d\mu_{|\ell}(y_1 \ldots y_\ell) \frac{\|\mu - \hat{\mu}_\ell\|_{\mathrm{TV}} \Delta^2 \gamma^{2\ell}}{(1-\gamma^2)\beta^2} = \frac{\|\mu - \hat{\mu}_\ell\|_{\mathrm{TV}} \Delta^2 \gamma^{2\ell}}{(1-\gamma^2)\beta^2} \ .\end{aligned}$$

Thus, $\|\mu - \hat{\mu}_\ell\|_{\beta,\gamma} \sim O(\beta^{-1}\gamma^\ell)$*.*

In summary, representing measures μ over $\mathcal{X}^{\mathbb{N}}$ by their truncated forms $\mu_{|\ell}$ leads to a Hilbert space representation that admits an approximate isomorphism to the space of full measures. The resulting truncation error is of order $O(\beta^{-1}\gamma^\ell)$.

We close this part with a minor note about a lower bound on the distance between measures. Given a word w, the function on $\mathcal{X}^\ell$ that equals 1 when $X = w$ and zero otherwise has a representation $|w\rangle^{(\ell)}_{\beta,\gamma}$ in $\mathcal{H}^{(\ell)}_{\beta,\gamma}$. (This follows since for finite $\mathcal{X}$, all functions on $\mathcal{X}^\ell$ belong to $\mathcal{F}^{(\ell)}_{\beta,\gamma}$.) The extension of this to $\mathcal{H}_{\beta,\gamma}$ is $|w\rangle_{\beta,\gamma} := |w\rangle^{(\ell)}_{\beta,\gamma} \otimes |\lambda_{\beta,\gamma}\rangle_{\beta\gamma^{-\ell},\gamma}$. This has the convenient property that $\langle w|\mu\rangle_{\beta,\gamma} = \Pr_\mu(\,w\,)$. Then, by the Cauchy-Schwarz inequality, for any measures μ and ν and any word w:

$$\begin{aligned}\|\mu - \nu\|_{\beta,\gamma} &\geq \frac{|\,\langle w|\mu - \nu\rangle\,|}{\sqrt{\langle w|w\rangle_{\beta,\gamma}}} \\ &= \frac{|\Pr_\mu(\,w\,) - \Pr_\nu(\,w\,)|}{\sqrt{\langle w|w\rangle_{\beta,\gamma}}} \ . \end{aligned} \tag{15}$$

So, word probabilities function as lower bounds on the Hilbert space norm.

D. Predictive states from kernel Bayes' rule

A prominent use of reproducing kernel Hilbert spaces is to approximate empirical measures [40]. Given a measure μ over a space $\mathcal{X}$ and N samples X_k drawn from this space, one constructs an approximate representation of μ via:

$$|\hat{\mu}\rangle := \frac{1}{N}\sum_{k=1}^{N} |\delta_{X_k}\rangle \ .$$

In other words, μ is approximated as a sum of delta functions centered on the observations. Convergence of this approximation to $|\mu\rangle$ is (almost surely) $O(N^{-1/2})$ [40].

This fact, combined with our Theorem 4, immediately gives the following result for $\mathcal{H}_{\beta,\gamma}$:

Proposition 6. *Suppose for some* $\mu \in \mathbb{P}(\mathcal{X}^{\mathbb{N}})$ *we take* N *samples of length* ℓ*, denoted* $\{X_k \in \mathcal{X}^\ell\}$ *(*$k = 1 \ldots N$*), and construct the state:*

$$|\hat{\mu}_{\ell,N}\rangle_{\beta,\gamma} = \frac{1}{N}\sum_{k=1}^{N} |\delta_{X_k}\rangle^{(\ell)}_{\beta,\gamma} \otimes |\lambda_{\beta\gamma^{-\ell},\gamma}\rangle_{\beta\gamma^{-\ell},\gamma} \ .$$

Then $|\hat{\mu}_{\ell,N}\rangle_{\beta,\gamma} \to |\mu\rangle$ *converges almost surely as* $N, \ell \to \infty$ *with error* $O(N^{-1/2} + \beta^{-1}\gamma^\ell)$*.*

Note the addition of the truncation error $O(\beta^{-1}\gamma^\ell)$ to the typical sample convergence $N^{-1/2}$. The truncation error has no dependence on the number of samples. It is a consequence of using an overly simplified hypothesis space $\mathcal{H}^{(\ell)}_{\beta,\gamma}$ to estimate μ.

A more nuanced application of RKHS for measures lies in reconstructing conditional distributions [26–28, 40, 41]. Let μ be a joint measure on some $\mathcal{X} \times \mathcal{Y}$, and let $\mu|_{\mathcal{X}}$ and $\mu|_{\mathcal{Y}}$ be its marginalizations. Given N samples (X_k, Y_k), construct the covariance operators:

$$\hat{C}_{XX} := \frac{1}{N}\sum_k |\delta_{X_k}\rangle\langle\delta_{X_k}| \ \text{ and}$$

$$\hat{C}_{YX} := \frac{1}{N}\sum_k |\delta_{Y_k}\rangle\langle\delta_{X_k}| \ .$$

Let $\mu_{\mathcal{Y}|X}$ be the conditional measure for $X \in \mathcal{X}$. For some $g \in \mathcal{H}_{\mathcal{Y}}$—the RKHS constructed on $\mathcal{Y}$—let $F_g(X) := \langle g|\mu_{\mathcal{Y}|X}\rangle$ be a function on $\mathcal{X}$. If $F_g \in \mathcal{H}_{\mathcal{X}}$ for all $g \in \mathcal{H}_{\mathcal{Y}}$, then $\hat{C}_{YX}\left(\hat{C}_{XX} - \zeta I\right)^{-1}|\delta_X\rangle$ converges to $|\mu_{\mathcal{Y}|X}\rangle$ as $N \to \infty$, $\zeta \to 0$, with convergence rate $O\left((N\zeta)^{-1/2} + \zeta^{1/2}\right)$.

The requirement essentially tells us that the structure of the conditional measure is compatible with the structures represented by the RKHS.

This is the *kernel Bayes' Rule* [41]. It applies to our $\mathcal{H}_{\beta,\gamma}$, by combining it with our results on truncated representations. For this theorem the reader should refer to Cors. 1 and 2. These define truncated predictive state $\eta_\ell[\overleftarrow{X}]$ and assert that it converges weakly to the predictive state $\epsilon[\overleftarrow{X}]$ for $\overleftarrow{\mu}$-almost all $\overleftarrow{X}$ as $\ell \to \infty$. This implies that

for any continuous function g;

$$\left\langle \eta_\ell[\overleftarrow{X}], g \right\rangle \to \left\langle \epsilon[\overleftarrow{X}], g \right\rangle \ ,$$

albeit at some unspecified rate $O(h_{\overleftarrow{X}}(\ell))$. (Here, we use $\langle \mu, f \rangle = \int f(X) d\mu(X)$, not to be confused with the inner products of $\mathcal{H}_{\beta,\gamma}$ and $\mathcal{F}_{\beta,\gamma}$).

Theorem 5. *Let $\mu \in \mathbb{P}(\mathcal{X}^{\mathbb{Z}})$ be a stationary and ergodic process. Suppose we take a long sample $X \in \mathcal{X}^L$ and from this sample subwords of length 2ℓ, $w_t = x_{t-\ell+1} \ldots w_{t+\ell}$ for $t = \ell, \ldots, L-\ell$. (There are $L - 2\ell + 1$ such words.) Split each word into a past $\overleftarrow{w}_t = x_{t-\ell+1} \ldots w_t$ and a future $\overrightarrow{w}_t = x_{t+1} \ldots x_{t+\ell}$, each of length ℓ. Define the operators:*

$$\hat{C}^{(\overleftarrow{X}\overleftarrow{X})}_{\beta,\gamma} = \frac{1}{L-2\ell+1} \sum_{t=\ell}^{L-\ell} |\hat{\delta}_{\overleftarrow{w}_t}\rangle_{\beta,\gamma} \otimes |\hat{\delta}_{\overleftarrow{w}_t}\rangle_{\beta,\gamma} \ \text{ and}$$

$$\hat{C}^{(\overleftarrow{X}\overrightarrow{X})}_{\beta,\gamma} = \frac{1}{L-2\ell+1} \sum_{t=\ell}^{L-\ell} |\hat{\delta}_{\overleftarrow{w}_t}\rangle_{\beta,\gamma} \otimes |\hat{\delta}_{\overrightarrow{w}_t}\rangle_{\beta,\gamma} \ .$$

Now, suppose for every $g \in \mathcal{F}_{\beta,\gamma}$ that $\left\langle \epsilon[\overleftarrow{X}], g \right\rangle \in \mathcal{F}_{\beta,\gamma}$ and $\left\langle \eta_\ell[\overleftarrow{X}], g \right\rangle \to \left\langle \epsilon[\overleftarrow{X}], g \right\rangle$ at a rate of $O(h_{\overleftarrow{X}}(\ell))$; see Section III E. Then for all $\overleftarrow{X}$:

$$\hat{C}^{(\overleftarrow{X}\overrightarrow{X})}_{\beta,\gamma} \left(\hat{C}^{(\overleftarrow{X}\overleftarrow{X})}_{\beta,\gamma} - \zeta \cdot I_{\beta,\gamma} \right)^{-1} |\delta_{\overleftarrow{X}}\rangle_{\beta,\gamma}$$

almost surely converges to $|\epsilon[\overleftarrow{X}]\rangle_{\beta,\gamma}$ as $L \to \infty$, $\ell \to \infty$, and $\zeta \to 0$, at the rate $O\left((L\zeta)^{-1/2} + \zeta^{1/2} + \gamma^{-\ell} + h_{\overleftarrow{X}}(\ell)\right)$.

This integrates all our results thus far with the usual kernel Bayes' rule. Several observations are in order.

1. First, there will ($\overleftarrow{\mu}$-almost) always be an $h_{\overleftarrow{X}}(\ell)$ as required by this theorem due to our own Cors. 1 and 2.

2. Second, since $\epsilon[\overleftarrow{X}]$ is not generally continuous, the theorem's strict requirements on $\epsilon[\overleftarrow{X}]$ are not satisfied. That said, weaker versions hold. If $\left\langle \epsilon[\overleftarrow{X}], g \right\rangle$ as a function of $\overleftarrow{X}$ does not belong to $\mathcal{F}_{\beta,\gamma}$ as a function of $\overleftarrow{X}$, then the representational error scaling depends on the precise form of $\epsilon[\overleftarrow{X}]$. From a learning theory perspective this error scaling depends upon the complexity of the hypothesis space relative to the function $\epsilon[\overleftarrow{X}]$ [42]. In the kernel embedding-of-distributions literature these scalings are obtained by choosing the ζ-parameter through cross-validation analysis [40, 41].

3. Third, all of this is contingent on our choice of metric $D_{\mathrm{E},\gamma}$ for the underlying sample space $\mathcal{X}^{\mathbb{Z}}$. The error whose scaling we have expressed above is formulated in terms of this metric, and so does not necessarily take into account whether the metric itself has been well-chosen. Supplementary Material II discusses the implications of the choice of metric.

V. EXAMPLES

Proposition 4 considered the existence of probabilistic expressions for the convergence rate of the truncated predictive states $\eta_\ell[\overleftarrow{X}]$ to the true predictive states $\epsilon[\overleftarrow{X}]$. We close with a handful of examples and case studies that give further insight to the convergence of $\left\| \eta_\ell[\overleftarrow{X}] - \epsilon[\overleftarrow{X}] \right\|_{\beta,\gamma}$ for widely-employed process classes—Markov, hidden Markov, and renewal processes.

A. Order-R Markov processes

A *Markov process* is a stochastic process where each observation x_t statistically depends only on the previous observation x_{t-1}. An order-R Markov process is one where each observation x_t depends only on the previous R observations $x_{t-R} \ldots x_{t-1}$. As such, the predictive states are simply given by:

$$\Pr_\mu \left(x \,\middle|\, \overleftarrow{X} \right) = \frac{\Pr_\mu (\, x_{-R+1} \ldots x_0 x \,)}{\Pr_\mu (\, x_{-R+1} \ldots x_0 \,)} \ ,$$

for each $\overleftarrow{X} = x_0 x_{-1} \ldots$. Since the predictive state is entirely defined after a finite number of observations, and this number is bounded by R, there is no conditioning error when R is taken as the observation length.

B. Hidden Markov processes

A *hidden Markov model* (HMM) $(\mathcal{S}, \mathcal{X}, \{ \mathbf{T}^{(x)} \})$ is defined here as a finite set $\mathcal{S}$ of states, a set $\mathcal{X}$ of observations, and a set $\mathbf{T}^{(x)} = (T^{(x)}_{ss'})$ of transition matrices, labeled by elements $x \in \mathcal{X}$ and whose components are indexed by $\mathcal{S}$ [21]. The elements are constrained so that $0 \leq T^{(x)}_{ss'} \leq 1$ and $\sum_{x,s'} T^{(x)}_{ss'} = 1$ for all $s, s' \in \mathcal{S}$. Let $\mathbf{T} = \sum_x \mathbf{T}^{(x)}$ and $\boldsymbol{\pi}$ be its left-eigenvector such that $\boldsymbol{\pi}\mathbf{T} = \boldsymbol{\pi}$. HMMs generate a stochastic process μ defined by the word probabilities:

$$\Pr_\mu (\, x_1 \ldots x_\ell \,) := \sum_{s'} \left[\boldsymbol{\pi} \mathbf{T}^{(x_1)} \ldots \mathbf{T}^{(x_\ell)} \right]_{s'} \ .$$

An extension of HMMs, called *generalized hidden Markov models* (GHMMs) [21] (or elsewhere *observable operator models* [22]), is defined as $(\mathbf{V}, \mathcal{X}, \{ \mathbf{T}^{(x)} \})$ where $\mathbf{V}$ is a finite-dimensional vector space. The only constraint on the transition matrices $\mathbf{T}^{(x)}$ is that $\mathbf{T}$ has a simple eigenvector of eigenvalue 1. The left-eigenvector is still denoted $\boldsymbol{\pi}$, the right-eigenvector denoted $\boldsymbol{\phi}$, and the word probabilities:

$$\Pr_\mu (x_1 \ldots x_\ell) := \boldsymbol{\pi} \mathbf{T}^{(x_1)} \ldots \mathbf{T}^{(x_\ell)} \boldsymbol{\phi}$$

are positive [21]. GHMMs generate a strictly broader class of processes than finite hidden Markov models can [21, 22, 43], though their basic structure is very similar.

First off, consider sofic processes. A *sofic process* is one that is not Markov at any finite order, but that is still expressible in a certain finite way. Namely, a sofic process is any that can be generated by a finite-state hidden Markov model with the *unifilar property*. An HMM has the unifilar property if $T^{(x)}_{s's} > 0$ only when $s' = f(x, s)$ for some deterministic function $f : \mathcal{S} \times \mathcal{X} \to \mathcal{S}$. Unifilar HMMs are the stochastic generalization of deterministic finite automata in computation theory [44].

The most useful property of sofic processes is that the states of their minimal unifilar HMM correspond exactly to the predictive states, of which there is always a finite number. Unlike with order-R Markov processes, there is no upper bound to how many observations it may take to δ-synchronize the predictive states. However, closed-form results on the synchronization to predictive states for unifilar HMMs is already known: at L past observations, with $L \to \infty$, the conditioning error is exponentially likely (in L) to be exponentially small (in L) [24]. In terms of our Hilbert space norm, there are constants α and C such that:

$$\Pr_{\overleftarrow{\mu}} \left(\left\| \eta_\ell[\overleftarrow{X}] - \epsilon[\overleftarrow{X}] \right\|_{\beta,\gamma} > \alpha^\ell \right) < C\alpha^\ell \ . \qquad (16)$$

As such, for $\overleftarrow{\mu}$-almost-all pasts, the corresponding convergence rate for the kernel Bayes' rule applied to a sofic process is $O\left((L\zeta)^{-1/2} + \zeta^{1/2} + \min(\alpha, \gamma)^{-\ell}\right)$.

Not all discrete-observation stochastic processes can be generated with a finite-state unifilar hidden Markov model. Though still encompassing only a small slice of processes, generalized hidden Markov models have a considerably larger scope of representation than finite unifilar models, as noted above.

The primary challenge in this setting is to relate the structure of a given HMM to the predictive states of its process. This is achieved through the notion of mixed states. A *mixed state* ρ is a distribution over the states of a finite HMM. A given HMM, with the stochastic dynamics between its own states, induces a higher-order dynamic on its mixed states and, critically for analysis, this is an *iterated function system* (IFS). Under suitable conditions the IFS has a unique invariant measure, and the support of this measure maps surjectively onto the process' set of predictive states. See Refs. [12] for details on this construction.

If $\rho = (\rho)$ is a mixed state, then the updated mixed state after observing symbol x is:

$$f^{(x)}_s(\rho) := \frac{1}{\sum_{s'} \left[\mathbf{T}^{(x)} \boldsymbol{\rho}\right]_{s'}} \left[\mathbf{T}^{(x)} \boldsymbol{\rho}\right]_s \ .$$

Let the matrix $\left[\mathbf{D} f^{(x)}\right]_{s's}(\rho)$ be given by the Jacobian $\partial f^{(x)}_{s'} / \partial \rho_s$ at a given value of ρ. There is a statistic, called the *Lyapunov characteristic exponent* $\lambda < 0$, such that:

$$\lambda = \lim_{\ell \to \infty} \frac{1}{\ell} \log \frac{\left\| \mathbf{D} f^{(x_\ell)}(\rho_\ell) \cdots \mathbf{D} f^{(x_1)}(\rho_1) \mathbf{v} \right\|}{\|\mathbf{v}\|} \ ,$$

where $\rho_t := f^{(x_{t-1})} \circ \cdots \circ f^{(x_1)}(\rho)$, for any vector $\mathbf{v}$ tangent to the simplex, almost any ρ (in the invariant measure), and almost any $\overrightarrow{X} = x_1 x_2 \ldots$ (in the measure of the prediction induced by ρ). The exponent λ then determines the rate at which conditioning error for predictive states converges to zero: for all δ and sufficiently large ℓ:

$$\Pr_{\overleftarrow{\mu}} \left(\left\| \eta_\ell[\overleftarrow{X}] - \epsilon[\overleftarrow{X}] \right\|_{\beta,\gamma} < Ce^{\lambda \ell} \right) > 1 - \delta \ .$$

This is somewhat less strict that Eq. (16)—depending on how rapidly the Lyapunov exponent converges in probability. In any case, for $\overleftarrow{\mu}$-almost all pasts, the convergence of the kernel Bayes' rule is $O\left((L\delta)^{-1/2} + \delta^{1/2} + \min(\lambda, \gamma)^{-\ell}\right)$, very similar to the sofic process rate.

We anticipate that these rules still broadly apply to generalized hidden Markov models, though we recommend more detailed analysis on this question.

C. Renewal processes

A renewal process, usually defined over continuous-time, can be defined for discrete time as follows. A renewal process emits 0s for a randomly selected duration before emitting a single 1 and then randomly selecting a new duration to fill with 0s [45]. Renewal processes can be as simple as Poisson processes, where the probability at any moment of producing another 0 or restarting on 1 is independent of time. Or, they can be far more memoryful, with a unique predictive state for any number of past 0s.

While high-memory renewal processes cannot generally

be represented by a finite hidden Markov model, they have only a countable number of predictive states, unlike most hidden Markov models. This follows since every number of past 0s defines a potential predictive state, but the process has no further memory beyond the most recent 1. Said simply, the predictive states are the time since last 1 [46]—or some coarse-graining of this indicator in special cases, such as the Poisson process.

A *renewal process* is specified by the survival probability $\Phi(n)$ that a contiguous block of 0s has length at least n. The exact probability of a given length is $F(n) := \Phi(n) - \Phi(n+1)$. It is always assumed that $\Phi(1) = 1$. Further, stationarity requires that $m := \sum_{n=1}^{\infty} \Phi(n)$ be finite, as this gives the mean length of a block of 0s. In the most general case the predictive states are given by:

$$\epsilon(\overleftarrow{X}) = \begin{cases} \epsilon_k & \overleftarrow{X} = 0^k 1 \dots \\ \text{undefined} & \overleftarrow{X} = 0^\infty \end{cases} ,$$

where the measures ϵ_k are recursively defined by the word probabilities:

$$\Pr_{\epsilon_k}\left(0^\ell 1 w \right) = \frac{F(k+\ell)}{\Phi(k)} \Pr_{\epsilon_0}\left(w \right) .$$

Now, it can be easily seen that each past $\overleftarrow{X}$ converges to zero conditioning error at a finite length since (almost) all pasts have the structure $\dots 10^k$, and so only the most recent $k+1$ values need be observed to know the predictive state. Therefore, the kernel Bayes' rule has an asymptotic convergence rate for each past $\overleftarrow{X}$ of $O\left((L\delta)^{-1/2} + \delta^{1/2} + \gamma^{-\ell}\right)$. However, this does not tell the entire story, as obviously not all pasts converge uniformly. A probabilistic expression of the conditioning error gives more information:

Proposition 7. *Suppose μ is a renewal process with $\Phi(n) \propto n^{-\alpha}$, $\alpha > 1$. Then there exist constants C and K such that:*

$$\Pr_{\overleftarrow{\mu}}\left(\left\| \eta_\ell(\overleftarrow{X}) - \epsilon(\overleftarrow{X}) \right\|_{\beta,\gamma} > C\ell^{-1} \right) > K\ell^{-\alpha} .$$

That is, the probability the conditioning error decays as $1/\ell$ is itself at least power-law decaying in ℓ.

Proof. *Recall from Eq.* (15)*:*

$$\left\| \eta_\ell(\overleftarrow{X}) - \epsilon(\overleftarrow{X}) \right\|_{\beta,\gamma} > \frac{\left| \Pr_\mu\left(w \mid x_1 \dots x_\ell \right) - \Pr_\mu\left(w \,\middle|\, \overleftarrow{X} \right) \right|}{\sqrt{\langle w|w \rangle_{\beta,\gamma}}} ,$$

for every word w, so we can choose any w and obtain a lower bound on the conditioning error. If our past $\overleftarrow{X}$ has the form $0^k 1 \dots$ for $k < \ell$, then we are already synchronized to the predictive state and the conditioning error is zero. Thus, we are specifically interested in the case $k \geq \ell$ and we will further consider the large-ℓ limit.

Now, under our assumptions, $\Phi(n) = n^{-\alpha}$ for some constant Z. For large n, $F(n) \sim \alpha n^{-\alpha-1}$. Then for any j:

$$\Pr_\mu\left(0^j 1 \,\middle|\, \overleftarrow{X} \right) = \frac{F(k+j)}{\Phi(k)} \sim \frac{\alpha}{k} \left(\frac{k+j}{k} \right)^{-\alpha-1} .$$

Meanwhile, so long as $k \geq \ell$, the truncated prediction has the form:

$$\begin{aligned} \Pr_\mu\left(0^j 1 \mid 0^\ell \right) &= \sum_{n=1}^{\infty} \frac{\Phi(n+\ell)}{\sum_p \Phi(p+\ell)} \frac{F(n+\ell+j)}{\Phi(n+\ell)} \\ &= \frac{\Phi(\ell+j)}{\sum_p \Phi(p+\ell)} \sim \frac{\alpha-1}{\ell} \left(\frac{\ell+j}{\ell} \right)^{-\alpha} . \end{aligned}$$

Now, choose $0 < C < \alpha - 1$ and define:

$$B = \left(1 - \frac{C+1}{\alpha} \right)^{-1} .$$

Then it can be checked straightforwardly that whenever $k > B\ell$, we have:

$$\begin{aligned} \Pr_\mu\left(1 \mid 0^\ell \right) - \Pr_\mu\left(1 \,\middle|\, \overleftarrow{X} \right) &\sim \frac{1}{\ell} \left[\alpha \left(1 - \frac{\ell}{k} \right) - 1 \right] \\ &> \frac{C}{\ell} . \end{aligned}$$

The probability that $k > B\ell$ is given by $\Phi(B\ell) = B^{-\alpha}\ell^{-\alpha}$. Setting $K = B^{-\alpha}/\sqrt{\langle 1|1 \rangle_{\beta,\gamma}}$ proves the theorem.

Therefore, while every sequence $\overleftarrow{X}$ converges to zero conditioning error at finite length, this convergence is not uniform, to such a degree that the proportion of pasts that retain conditioning error of $1/\ell$ has a fat tail in ℓ. This is a matter of practical importance that is not cleanly expressed in the big-O expression of the conditioning error from Thm. 5.

Poisson and renewal processes are merely the first two steps in a structural hierarchy of increasing sophistication. The next generalization beyond renewal processes are the *semi-Markov processes* and beyond those, the *hidden semi-Markov processes*. Roughly speaking, these are finite-state-controlled renewal process and moving up the hierarchy requires using more memoryful controllers. In this way, each process class in the hierarchy is expo-

nentially larger than it predecessor and so exponentially more expressive of complex process organization. Reference [45]'s results on hidden semi-Markov processes allow one to identify their predictive states. This section made the route to these generalizations explicit and so we do not include them here. We now shift to consider another process class that employs infinite memory in a different, more syntactic way than renewal process "count up and reset".

D. Aperiodic Infinitary Processes

Substitution shifts provide an easy-to-grasp but behaviorally nontrivial exploration ground for the possible behaviors of highly complex processes that are still ergodic. The Thue-Morse process, for instance, can be generated by starting from the string "0" and taking the limit of an infinite number of substitutions of the form:

$$\begin{aligned} 0 &\mapsto 01 \\ 1 &\mapsto 10 \ , \end{aligned}$$

and randomly sampling any contiguous block from the result. Similarly, using the substitution rule:

$$\begin{aligned} 0 &\mapsto 01 \\ 1 &\mapsto 10 \ , \end{aligned}$$

generates the Feigenbaum process. This has a physical interpretation. If we sample a sequence of values y_t from the logistic map:

$$y_{t+1} = r y_t (1 - y_t)$$

with $y_0 \in [0, 1]$ at critical parameter $r \approx 3.56995$, and only measure the function:

$$x_t(y_t) = \begin{cases} 0 & 0 \leq y_t < 1/2 \\ 1 & 1/2 \leq y_t \leq 1 \end{cases} \ ,$$

then the resulting sequence $X = (x_t)$ has the same statistics as the Feigenbaum process [29, 47].

These processes are considered *aperiodic* since they never fully repeat and *infinitary* since many measures of their long-time statistical dependencies diverge, such as correlation length and excess entropy [48, 49]. These processes are highly non-Markovian and sequential generation requires a nested stack (infinite memory) automaton [47, 48].

Nonetheless, our approach to predictive states here applies directly to these processes. Helpfully, these processes provide a useful example of the saturation of our convergence results. While the preceding focused on demonstrating that predictive states converge in distribution, for substitution shifts it can be seen that predictive states *fail* to converge under certain stronger criteria.

Thue-Morse and Feigenbaum processes are considered deterministic since they have an asymptotically vanishing entropy rate [48] and the nested-stack automaton controller is deterministic in the sense of formal language theory [44]. This means, in particular, that the probability mass in $\Pr_\mu\left(x_1 \,\middle|\, \overleftarrow{X} \right)$ is always concentrated on a 0 or 1. This can be extended to the observation that $\epsilon[\overleftarrow{X}]$, for $\overleftarrow{\mu}$-almost all $\overleftarrow{X}$, is a δ-measure concentrated on some specific future $\overrightarrow{X} = F(\overleftarrow{X})$.

A somewhat stronger criterion for convergence of measures is convergence over sets; that is, $\mu_n(A) \to \mu(A)$ for any measurable set A. An even stronger criterion is convergence in total variation: $\|\mu_n - \mu\|_{\mathrm{TV}} \to 0$.

However, suppose we choose the set $A = \left\{ F(\overleftarrow{X}) \right\}$, where F is the deterministic mapping between pasts and futures from before. Due to aperiodicity, $\eta_\ell[\overleftarrow{X}]$ cannot be concentrated on $F(\overleftarrow{X})$. It must be, in fact, diffuse. If this were not so, then there would be a word of infinite length—namely, $x_{-\ell+1} \ldots x_0 F(\overleftarrow{X})$—with nonzero measure under μ. This word must also be aperiodic, since this is a key property of samples generated by substitution shifts. Stationarity then implies that every time-shifted version of the word also carries nonzero probability. Aperiodicity then implies an infinite number of these. Whence we have a contradiction, since μ is a probability distribution and can only assign a finite total mass.

Consequently, $\eta_\ell[\overleftarrow{X}]$ must be a diffuse measure with no singular points, so $\eta_\ell[\overleftarrow{X}](A) = 0$ for all ℓ. Since $\epsilon[\overleftarrow{X}](A) = 1$, both convergence over sets and convergence in total variation fail for the predictive states of aperiodic infinitary processes. The very practical lesson is that attempts to recover such processes from empirical data *must* use tools that rely on convergence in *distribution*, such as the RKHS methods Sec. IV outlined.

VI. CONCLUDING REMARKS

Taken altogether, the results fill-in important gaps in the foundations of predictive states, while strengthening those foundations for further development, extension, and application. Previously, the properties of predictive states were only examined in the context of hidden Markov models, their generalizations, and hidden semi-Markov models. We provided a definition applicable to any stationary and ergodic process with discrete

and real-valued observations, extending previous foundational work [21]. Further, we showed that predictive states for all such processes are learnable from empirical data, whether through a direct method of partitioning pasts or through indirect methods, such as the reproducing kernel Hilbert space.

One important extension is to continuous-time processes. By exploiting the full generality of Jessen's and Enomoto's theorems we believe this extension is quite feasible. As long as the set of possible pasts and futures constitutes a separable space, they should be expressible in the form of a countable basis, to which these theorems may then be applied. (See our example in Supplementary Material I.) The challenge lies in constructing an appropriate and useful basis. We leave this for future work.

We described key properties of the space in which predictive states live. However, predictive states are not merely static objects. They predict the probabilities of future observations. And, once those observations are made, the predictive state may be updated to account for new information. Thus, predictive states provide the stochastic rules for their own transformation into future predictive states. This dynamical process has been explored in great detail in the cases where the process is generated by a finite hidden Markov model—this is found in former work on the ϵ-machine and the mixed states of HMMs. (See, for instance, Refs. [12, 50, 51].) Understanding the nature of this dynamic for more general processes, including how it makes contact with other dynamical approaches like stochastic differential equations in the continuous-time setting, also remains for future work.

ACKNOWLEDGMENTS

We thank Nicolas Brodu and Adam Rupe for helpful comments and revisions. We also thank Alex Jurgens, Greg Wimsatt, Fabio Anza, David Gier, Kyle Ray, Mikhael Semaan, and Ariadna Venegas-Li for helpful discussions. JPC acknowledges the kind hospitality of the Telluride Science Research Center, Santa Fe Institute, Institute for Advanced Study at the University of Amsterdam, and California Institute of Technology for their hospitality during visits. This material is based upon work supported by, or in part by, Grant Nos. FQXi-RFP-IPW-1902 and FQXi-RFP-1809 from the Foundational Questions Institute and Fetzer Franklin Fund (a donor-advised fund of Silicon Valley Community Foundation), grant TWCF0570 from the Templeton World Charity Foundation, grants W911NF-18-1-0028 and W911NF-21-1-0048 from the U.S. Army Research Laboratory and the U.S. Army Research Office, and grant DE-SC0017324 from the U.S. Department of Energy.

[1] J. P. Crutchfield. Between order and chaos. *Nature Physics*, 8(January):17–24, 2012. 1

[2] A. B. Boyd and J. P. Crutchfield. Maxwell demon dynamics: Deterministic chaos, the Szilard map, and the intelligence of thermodynamic systems. *Phys. Rev. Let.*, 116:190601, 2016. 1

[3] S. Loomis and J. P. Crutchfield. Thermal efficiency of quantum memory compression. *Phys. Rev. Lett.*, 125:020601, 2020.

[4] A. Jurgens and J. P. Crutchfield. Functional thermodynamics of maxwellian ratchets: Constructing and deconstructing patterns, randomizing and derandomizing behaviors. *Phys. Rev. Res.*, 2(3):033334, 2020. 1

[5] M. Gu, K. Wiesner, E. Rieper, and V. Vedral. Quantum mechanics can reduce the complexity of classical models. *Nature Comm.*, 3(762):1–5, 2012. 1

[6] J. R. Mahoney, C. Aghamohammadi, and J. P. Crutchfield. Occam's quantum strop: Synchronizing and compressing classical cryptic processes via a quantum channel. *Scientific Reports*, 6:20495, 2016.

[7] F. C. Binder, J. Thompson, and M. Gu. A practical, unitary simulator for non-Markovian complex processes. arXiv.org:1709.02375.

[8] A. Venegas-Li, A. Jurgens, and J. P. Crutchfield. Measurement-induced randomness and structure in controlled qubit processes. *Phys. Rev. E*, 102(4):040102(R), 2020. 1

[9] J. P. Crutchfield and D. P. Feldman. Statistical complexity of simple one-dimensional spin systems. *Phys. Rev. E*, 55(2):R1239–R1243, 1997. 1

[10] D. P. Varn and J. P. Crutchfield. Chaotic crystallography: How the physics of information reveals structural order in materials. *Curr. Opin. Chem. Eng.*, 7:47–56, 2015.

[11] S. E. Marzen and J. P. Crutchfield. Optimized bacteria are environmental prediction engines. *Phys. Rev. E*, 98:012408, 2018. 1

[12] A. Jurgens and J. P. Crutchfield. Divergent predictive states: The statistical complexity dimension of stationary, ergodic hidden Markov processes. *Chaos*, 2021. 1, 12, 17, 20

[13] A. Rupe and J. P. Crutchfield. Local causal states and discrete coherent structures. *Chaos*, 28(7):1–22, 2018. 1

[14] C. R. Shalizi, K. L. Shalizi, and J. P. Crutchfield. Pattern discovery in time series, Part I: Theory, algorithm, analysis, and convergence. *arXiv.org/abs/cs.LG/0210025*. 1

[15] S. Still, J. P. Crutchfield, and C. J. Ellison. Optimal causal inference: Estimating stored information and approximating causal architecture. *CHAOS*, 20(3):037111, 2010.

[16] C. C. Strelioff and J. P. Crutchfield. Bayesian structural inference for hidden processes. *Phys. Rev. E*, 89:042119, 2014.

[17] S. Marzen and J. P. Crutchfield. Predictive rate-distortion for infinite-order markov processes. *J. Stat. Phys.*, 163(6):1312–1338, 2014.

[18] A. Rupe, N. Kumar, V. Epifanov, K. Kashinath, O. Pavlyk, F. Schimbach, M. Patwary, S. Maidanov, V. Lee, Prabhat, and J. P. Crutchfield. Disco: Physics-based unsupervised discovery of coherent structures in spatiotemporal systems. In *2019 IEEE/ACM Workshop on Machine Learning in High Performance Computing Environments (MLHPC)*, pages 75–87, 2019.

[19] A. Rupe and J. P. Crutchfield. Spacetime autoencoders using local causal states. *AAAI Fall Series 2020 Symposium on Physics-guided AI for Accelerating Scientific Discovery*, 2020. arXiv:2010.05451.

[20] N. Brodu and J. P. Crutchfield. Discovering causal structure with reproducing-kernel Hilbert space ϵ-machines. *arXiv:2011.14821*, 2020. 1

[21] D. R. Upper. *Theory and Algorithms for Hidden Markov Models and Generalized Hidden Markov Models*. PhD thesis, University of California, Berkeley, 1997. Published by University Microfilms Intl, Ann Arbor, Michigan. 1, 4, 16, 17, 20

[22] H. Jaeger. Observable operator models for discrete stochastic time series. *Neural Computation*, 12(6):1371–1398, 2000. 17

[23] L. Song, B. Boots, S. Siddiqi, G. Gordon, and A. Smola. Learning and discovery of predictive state representations in dynamical systems with reset. In *Proceedings of the Twenty-first International Conference on Machine Learning*, page 53. ACM, 2004.

[24] N. Travers and J. P. Crutchfield. Asymptotic synchronization for finite-state sources. *J. Stat. Phys.*, 145(5):1202–1223, 2011. 17

[25] M. Thon and H. Jaeger. Links between multiplicity automata, observable operator models and predictive state representations – a unified learning framework. *J. Mach. Learn. Res.*, 16:103–147, 2015. 1

[26] L. Song, J. Huang, A. Smola, and K. Fukumizu. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *Proceedings of the 26th International Conference on Machine Learning*, page 961–968. ACM, 2009. 1, 15

[27] L. Song, B. Boots, S. Siddiqi, G. Gordon, and A. Smola. Hilbert space embeddings of hidden markov models. In *Proceedings of the 27th International Conference on Machine Learning*, pages 991–998. Omnipress, 2010.

[28] B. Boots, A. Gretton, and G. Gordon. Hilbert space embeddings of predictive state representations. In *Proceedings of the 29th International Conference on Uncertainty in Artificial Intelligence*, pages 92–101, 2013. 1, 15

[29] O. Kallenberg. *Foundations of Modern Probability*. Springer, New York, 2 edition, 2001. 2, 3, 19

[30] P. Kůrka. *Topological and Symbolic Dynamics*. Société Mathématique de France, Paris, 2003. 2

[31] M. M. Rao. *Conditional Measures and Applications*. CRC Press, Boca Raton, FL, second edition, 2005. 4, 5, 6

[32] B. Jessen. The theory of integration in a space of an infinite number of dimensions. *Acta Math.*, 63:249–323, 1934. 6, 7, 8, 10

[33] B. Jessen. A remark on strong differentiation in a space of infinitely many dimensions. *Mat. Tidsskr. B.*, pages 54–57, 1952. 6, 10

[34] P. S. Enomoto. Dérivation par rapport à un système de voisinages dans l'espace de tore. *Proc. Japan Acad.*, 30(8):721–725, 1954. 6, 10

[35] N. Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 68:337–404, 1950. 11

[36] B. Sriperembudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. R. G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *J. Machine Learn. Res.*, 11:1517–1561, 2010. 11, 13

[37] I. Steinwart. Reproducing kernel Hilbert spaces cannot contain all continuous functions on a compact metric space. 2020. arXiv: 2002.03171. 11

[38] A. Christman and I. Steinwart. Universal kernels on non-standard input spaces. In *Proc. 23rd Intl. Conf. Neural Information Processing Systems*, volume 1, pages 406–414. ACM, 2010. 13, 1, 4

[39] H. Zhang and L. Zhao. On the inclusion relation of reproducing kernel Hilbert spaces. *Analysis and Applications*, 11(2):1350015, 2013. 13

[40] K. Muandet, K. Fukumizu, B. Sriperumbudur, and B. Schölkopf. Kernel mean embedding of distributions: A review and beyond. *Found. Trends in Machine Learn.*, 10(1-2):1–141, 2017. 15, 16

[41] K. Fukumizu, L. Song, and A. Gretton. Kernel bayes' rule: Bayesian inference with positive definite kernels. *J. Mach. Learn. Res.*, 14:3753–3783, 2013. 15, 16

[42] F. Cucker and D. X. Zhou. *Learning Theory: An Approximation Theory Viewpoint*. Cambridge University Press, Cambridge, 2007. 16

[43] H. Ito, S.-I. Amari, and K. Kobayashi. Identifiability of hidden Markov information sources and their minimum degrees of freedom. *IEEE Info. Th.*, 38:324, 1992. 17

[44] J. E. Hopcroft, R. Motwani, and J. D. Ullman. *Introduction to Automata Theory, Languages, and Computation*. Prentice-Hall, New York, third edition, 2006. 17, 19

[45] S. Marzen and J. P. Crutchfield. Structure and randomness of continuous-time discrete-event processes. *J. Stat. Physics*, 169(2):303–315, 2017. 17, 19

[46] S. Marzen and J. P. Crutchfield. Informational and causal architecture of continuous-time renewal processes. *J. Stat. Phys.*, 168(a):109–127, 2017. 18

[47] J. P. Crutchfield. The calculi of emergence: Computation, dynamics, and induction. *Physica D*, 75:11–54, 1994. 19, 5

[48] K. Young. *The Grammar and Statistical Mechanics of Complex Physical Systems*. PhD thesis, University of

California, Santa Cruz, 1991. published by University Microfilms Intl, Ann Arbor, Michigan. 19

[49] J. P. Crutchfield and D. P. Feldman. Regularities unseen, randomness observed: Levels of entropy convergence. *CHAOS*, 13(1):25–54, 2003. 19, 4, 5

[50] A. Jurgens and J. P. Crutchfield. Shannon entropy rate of hidden Markov processes. *J. Statistical Physics*, 183(32):1–18, 2020. 20

[51] A. Jurgens and J. P. Crutchfield. Ambiguity rate of hidden Markov processes. *arxiv.org:2105.07294*, 2020. 20

[52] A. Venegas-Li and J. P. Crutchfield. Optimality and complexity in measured quantum-state stochastic processes. 2022. arXiv:2205.03958 [quant-ph]. 1

[53] P. R. Halmos. *A Hilbert Space Problem Book*. Springer, New York, NY, second edition, 1982. 2

Supplementary Materials

Topology, Convergence, and Reconstruction of Predictive States

Samuel P. Loomis and James P. Crutchfield

The Supplementary Materials consider extensions and open questions beyond the main body's scope, but that nonetheless may be of interest to the reader.

I. PREDICTIVE STATES FOR MORE GENERAL STOCHASTIC PROCESSES

The main text considered the case where $\mathcal{T} = \mathbb{Z}$ and where $\mathcal{X}$ is compact. Compactness, in particular, is significant in two places: It underlies Prop. 2 that relates predictive-state convergence to convergence in distribution. Additionally, Ref. [38]'s theorem assumes it to guarantee that the Gaussian-kernel RKHS metrizes the topology of convergence in distribution.

The following discusses various ways to relax these assumptions and so how Theorem 2 may still be used to define and characterize predictive-state convergence in much broader settings, including continuous-time processes.

A. Relaxing (and reimposing) compactness

The most basic relaxation of compactness is to allow $\mathcal{X}$ to have infinite range. For instance, one way to characterize point processes—such as, neural spike trains—is as a sequence of interevent times $(t_1, t_2, \dots)$ that, depending on the setting, may be arbitrarily large. In this case, $\mathcal{X} = [0, \infty)$.

A straightforward way to topologically address this case is to simply compactify $\mathcal{X}$ with the addition of a point at infinity. However, this impacts the choice of distance metric on $\mathcal{X}$ used to construct $D_{\mathrm{E},\gamma}$. The distance metric must metrize the compactified topology. One option for this is:

$$d(t_1, t_2) = |\tanh(t_2) - \tanh(t_1)|$$

This is simply the metric obtained by processing the data $t \mapsto x$ under the rule $x(t) = \tanh(t)$. Any such metric must de-emphasize differences between large values relative to equal-magnitude differences between small values. This method works because $\mathbb{R}$ is locally compact and, therefore, has a one-point compactification. This method would similarly extend to any locally compact symbol space.

A less straightforward case is when $\mathcal{X}$ takes values in an infinite-dimensional Banach or Hilbert space. This may occur, for instance, if the stochastic process symbols are quantum-state valued, existing on the unit sphere $S_{\mathcal{H}}$ in a Hilbert space $\mathcal{H}$, as in Refs. [8, 52]. If $\mathcal{H}$ is infinite-dimensional, the unit sphere $S_{\mathcal{H}}$ is compact in the weak topology of $\mathcal{H}$ but not the norm topology. Measures over $S_{\mathcal{H}}$ (or its subsets) may be endowed with the topology of convergence in distribution with respect to *either* the weak or norm topology of $\mathcal{H}$. The considerations below suggest that predictive states can be conditioned on the past in a manner that converges in distribution in *either sense*. However, it is only in the compact weak topology of $S_{\mathcal{H}}$ that we can assume a Gaussian RKHS metrizes the topology of convergence in distribution. This affects our choice of measure.

Suppose we take a sequence of quantum states $\overleftarrow{\Psi} = (|\psi_{-t}\rangle)_{t \in \mathbb{N}}$ and break them down according to some orthogonal basis $|e_j\rangle$:

$$|\psi_{-t}\rangle = \sum_{j=1}^{\infty} c_{t,j} \, |e_j\rangle \ .$$

(This assumes $\mathcal{H}$ is separable, but in physical settings this is generally taken for granted.) We enumerate the coefficients

into a *single sequence* according to a zig-zag scheme:

$$\begin{pmatrix} c_{1,1} & c_{1,2} & c_{1,3} & \cdots \\ c_{2,1} & c_{2,2} & & \\ & & \ddots & \\ c_{3,1} & & & \\ \vdots & & & \end{pmatrix} \mapsto (C_j) := (c_{1,1}, c_{2,1}, c_{1,2}, c_{1,3}, c_{2,2}, c_{3,1}, \dots) \ .$$

This scheme allows enumeration to explore both any past time or any coefficient depth in a finite number of steps. It has the effect of pushing steadily further into the past while doubling back to add further detail to the more recent quantum states. Given any neighborhood U in $S_{\mathcal{H}}$, in whichever chosen topology, we define the predictive-state probability as:

$$\Pr_\mu \left(U \,\middle|\, \overleftarrow{\Psi} \right) = \lim_{j \to \infty} \Pr_\mu (U \mid C_1, \dots, C_j) \ .$$

This conditioning scheme depends only on a countable, bounded sequence of complex numbers; that may be seen as isomorphic to $\mathbb{R}^2$. It ultimately contains all information about past states. And, to the point, it is convergent as a consequence of the generalized Enomoto theorem Thm. 2. Since separable Hilbert spaces are second-countable, we need only collect these probabilities on a countable number of neighborhoods U to uniquely determine the predictive state on all sets. This guarantees that there exists a measure-one set of pasts on which the predictive state converges in distribution.

As noted, the results on RKHS representations of predictive states only hold if the underlying observation space is compact. To construct a Gaussian RKHS over $S_{\mathcal{H}}$ we must metrize its weak topology. As it happens, the weak topology is in general not metrizable on infinite-dimensional $\mathcal{H}$, but it *is* metrizable on any bounded subset [53], including $S_{\mathcal{H}}$. The simplest such metric, perhaps unsurprisingly, has the form:

$$d_\gamma(\psi_1, \psi_2)^2 := \sum_{j=1}^{\infty} \gamma^{2j} \left| c_{1,j} - c_{2,j} \right|^2$$

for $0 < \gamma < 1$. And so, we may define:

$$\begin{aligned} D_{\mathrm{E},\gamma_1,\gamma_2}(\Psi, \Phi) &:= \sum_{t=1}^{\infty} \gamma_2^{2t} d_{\gamma_1}(\psi_t, \phi_t)^2 \\ &= \sum_{t,j} \gamma_2^{2t} \gamma_1^{2j} \left| c_{\psi_t,j} - c_{\phi_t,j} \right|^2 \end{aligned}$$

as the Hilbert-space analogue of $D_{\mathrm{E},\gamma}$. This, rather than a norm-based approach, is the most appropriate metric for constructing a Gaussian RKHS for inference on a sequence of infinite-dimensional states.

To summarize, the simplest useful approach to acquire predictive states over noncompact observation spaces is to make them compact, either by compactification (for unbounded observations) or adopting the weak topology (for Banach/Hilbert spaces).

B. Continuous time processes

If $\mathcal{T} = \mathbb{R}$ we open ourselves up to a much richer world of stochastic processes. The data now becomes functions from $\mathbb{R}$ to $\mathbb{R}$. This forces adopting more care about defining the full measure space. The following sketches out the broad lines, but leaves formal development for future effort.

For simplicity, we constrain ourselves to consider only right-continuous data with left limits; that is, *càdlàg* functions. The first thing to note in this case is that right-continuity means we can fully characterize any càdlàg function X by knowing its values over only a countably dense subset $\mathcal{D} \subseteq \mathbb{R}$. As in the case of the Hilbert space sequences, we can concoct a well-ordered enumeration (t_j) of $\mathcal{D}$ that gradually explores more distant pasts while also incorporating

more frequent recent observations. For instance, suppose we take $\mathcal{D} \subset [0, \infty)$ defined by:

$$\mathcal{D} = \left\{ \frac{n}{2^j} : n \in \mathbb{N}, j \in \mathbb{N} \right\} .$$

We may take the enumeration:

$$(t_j) := \left(0, 1, \frac{1}{2}, \frac{1}{4}, \frac{3}{4}, \frac{3}{2}, 2, \ldots \right) .$$

This follows the same zigzag scheme of our Hilbert space example. Using this enumeration of past times, we may define the limit:

$$\Pr\left(U \,\middle|\, \overleftarrow{X} \right) = \lim_{j \to \infty} \Pr\left(U \mid X(-t_1), \ldots, X(-t_j) \right) ,$$

where X is a function on $\mathbb{R}$ and $\overleftarrow{X}$ represents its projection to $(-\infty, 0)$. Again, since this is a countable sequence of real numbers, we may apply the generalized Enomoto theorem Thm. 2 to conclude convergence for $\overleftarrow{\mu}$-almost all $\overleftarrow{X}$.

What kind of neighborhood U should we use in the above limit? There are a few different ways of defining a topology on càdlàg functions. A popular choice is the Skorohod metric—not defined here since it is not used—that essentially seeks the regularized "supremum distance" between two càdlàg functions optimized over a relative "wiggling" of the time variable. We propose instead to use the following L^2-norm for functions X, Y over $[0, \infty)$:

$$\|X - Y\|_\Gamma^2 = \int_0^\infty |X(t) - Y(t)|^2 \, d\Gamma(t) ,$$

where Γ is any probability measure over $[0, \infty)$.

For instance, if we desire an analogy with $D_{\mathrm{E},\gamma}$ in the discrete case, we may take $d\Gamma(t) = \gamma^{2t} dt$ for some $0 < \gamma < 1$. Compact sets in the $L^2(\Gamma)$ norm are characterized by a form of the Kolmogorov-Riesz theorem: a set $\overline{\mathcal{X}} \subset \mathcal{X}^{[0,\infty)}$ is compact if and only if:

1. $\overline{\mathcal{X}}$ is bounded,
2. for every $\delta > 0$ there is $T > 0$ so that for every $X \in \overline{\mathcal{X}}$:

$$\int_T^\infty |X(t)|^2 \, d\Gamma(t) < \delta$$

   and,

3. for every $\delta > 0$ there is $\Delta > 0$ so that, for every $X \in \overline{\mathcal{X}}$ and $\tau < \Delta$,

$$\int_0^\infty |X(t + \tau) - X(t)|^2 \, d\Gamma(t) < \delta$$

   .

The first criterion ensures that X cannot not have arbitrarily large values at the origin. The second ensures that X cannot grow "too fast". The third ensures that X does not vary "too fast". In other words, if we assume the continuous signal X to be totally bounded in magnitude and to be somewhat rate-limited in the frequency of its variation, then the range of possible X values will be compact in $L^2(\mu)$. Furthermore, the norm $\|\cdot\|_\Gamma$ is decomposable over time in a similar way to $D_{\mathrm{E},\gamma}$, and if we adopt $d\Gamma(t) = \gamma^{2t} dt$, then it obeys an analogous Pythagorean theorem. This should make extending Theorems 4 and 5 relatively straightforward.

II. METRIZING PRODUCT TOPOLOGY

The product topology of $\mathcal{X}^{\mathbb{N}}$ is metrizable—a fact use in the main development. The metric used depends on a parameter γ:

$$D_{\mathrm{E},\gamma}(X,Y)^2 := \begin{cases} \sum_{t=1}^{\infty}(1-\delta_{x_t y_t})\gamma^{2t} & \mathcal{X} \text{ discrete} \\ \sum_{t=1}^{\infty} \|x_t - y_t\|^2 \gamma^{2t} & \mathcal{X} \subset \mathbb{R}^d \end{cases} .$$

The main development's primary concern is only to relate the topology governing convergence of predictive states to the natural topology of RKHS's. In particular, a Gaussian kernel RKHS defined over a compact region of a Hilbert space reproduces the topology of convergence in distribution [38]. The main development showed that $D_{\mathrm{E},\gamma}$ can be interpreted as arising from embedding $\mathcal{X}^{\mathbb{N}}$ in a Hilbert space, and so the Gaussian kernel RKHS constructed from this metric preserves the convergence of predictive states.

However, convergent predictive-state approximation says little about convergence speed. The asymptotic error rates in Theorem 5 are extremely open-ended, and their coefficients may depend nontrivially on the choice of how we metrize $\mathcal{X}^{\mathbb{N}}$. Determining this precise dependence is beyond the scope of the present development, but the following still offers several brief caveats on metric choice.

First, we note that while the $D_{\mathrm{E},\gamma}$ for $0 < \gamma < 1$ are all topologically equivalent, they are not metrically equivalent. That is, there are no constants $\alpha, \beta > 0$ such that:

$$\alpha D_{\mathrm{E},\gamma_2}(X,Y) \leq D_{\mathrm{E},\gamma_1}(X,Y) \leq \beta D_{\mathrm{E},\gamma_2}(X,Y) ,$$

for all $X, Y \in \mathcal{X}^{\mathbb{N}}$ when $\gamma_2 \neq \gamma_1$. Specifically, suppose that $\gamma_2 > \gamma_1$. Given some α, choose an integer K such that:

$$K > -\frac{\log(\alpha)}{\log(\gamma_2/\gamma_1)} .$$

Choose any X and Y such that $x_t = y_t$ for all $1 \leq t < K$ and $x_K \neq y_K$. Then:

$$\begin{aligned} D_{\mathrm{E},\gamma_2}(X,Y)^2 &= \sum_{t=K}^{\infty} \gamma_2^{2K}(1-\delta_{x_k y_k}) \\ &= \sum_{t=K}^{\infty} \gamma_1^{2K} \left(\frac{\gamma_2}{\gamma_1}\right)^{2K} (1-\delta_{x_k y_k}) \\ &> \frac{1}{\alpha^2} D_{\mathrm{E},\gamma_1}(X,Y)^2 . \end{aligned}$$

And so, D_{E,γ_2} and D_{E,γ_1} are not metrically equivalent. This means the choice of γ has a meaningful consequence on the geometry of the processed data and on the reconstructed predictive states.

Additionally, while the development intentionally chose the $D_{\mathrm{E},\gamma}$ for simplicity, it need not have restricted to exponential decay. Alternative metrizations of the product topology exist. The most straightforward extension would be of the form:

$$D_{\mathrm{E},W}(X,Y)^2 := \begin{cases} \sum_{t=1}^{\infty}(1-\delta_{x_t y_t})W(t)^2 & \mathcal{X} \text{ discrete} \\ \sum_{t=1}^{\infty} \|x_t - y_t\|^2 W(t)^2 & \mathcal{X} \subset \mathbb{R}^d \end{cases} ,$$

where $W(t) > 0$ is a weight function with finite squared sum $\sum_{t=1}^{\infty} W(t)^2 < \infty$. One could, for instance, adopt a power-law weight decay $W(t) = 1/t^{p/2}$, $p > 1$, giving a zeta-function distance metric.

We conjecture that the most convenient choice would tie the weight function $W(t)$ to a relevant time-scale of the stochastic process in question. We propose the following heuristic. Inspired by Ref. [49], define the *mutual information*

between the past $\overleftarrow{X}$ and the next ℓ symbols $x_1 \ldots x_\ell$ as:

$$I(\ell) = \int d\overleftarrow{\mu}(\overleftarrow{X}) \int d\mu_{|\ell}(x_1 \ldots x_\ell) \log \left(\frac{d\epsilon[\overleftarrow{X}]_{|\ell}}{d\mu_{|\ell}}(x_1 \ldots x_\ell) \right) \ .$$

This captures, essentially, the *redundancy* of information in the process. What information (of the most recent observations) did we already know from the past? If it is low, then future observations do not depend much on the past, and so temporal correlation is limited. From this, we define the *redundancy per symbol* at ℓ as:

$$r(\ell) = I(\ell) - I(\ell - 1) \ ,$$

supposing that $I(0) = 0$. This captures the extent to which an observation ℓ steps in the future exhibits *new correlations* with the past not captured by intermediate observations. We suggest that $W(t) \sim O(r(t))$ is an appropriate weight scheme for the distance measurement, as it weights the observation at each time by the potentially new dependence which that observation may have on the past.

In the discrete case, these quantities are readily estimated from available data, providing a rough idea of the weights to use. We provide a few examples, drawn from Ref. [49], that demonstrate how this weight scheme works:

1. For Markov processes of order R (including periodic processes of period R), $r(t) = 0$ for $t > R$. The distance metric $D_{\mathrm{E},W}$ therefore only compares the first R symbols of any sequence—the only symbols that directly depend upon the past. This saves computation time and reduces the effective complexity of the reproducing kernel Hilbert space.
2. For hidden Markov models, the redundancy per symbol follows an exponential scaling $r(t) \propto \gamma^t$ for some $0 < \gamma < 1$. In this case, the original choice in the main development—$D_{\mathrm{E},\gamma}$—is appropriate, with the parameter γ determined by $r(t)$.
3. To consider a fully non-Markovian process, we examine the Thue-Morse sequence, generated by the substitution rules $0 \mapsto 01$, $1 \mapsto 10$. Thue-Morse process cannot be generated by any hidden Markov model. It, in fact, corresponds to an indexed grammar [47]. It is also infinitary, meaning that $I(\ell)$ diverges as $\ell \to \infty$. Specifically, $I(\ell) \sim \log \ell$; then $r(\ell) \sim O(\ell^{-1})$. Choosing $W(\ell) = \ell^{-1}$ gives us a distance-squared sum that scales as ℓ^{-2} and so is still convergent.

To summarize, the weight that the metric over sequences assigns to future observations should depend on the process' estimated temporal correlations. We believe the most useful correlation estimate is in the redundancy per observation $r(\ell)$.